This International Student Edition is for use outside of the U.S.

Introduction to BUSINESS ANALYTICS

VERNON J. RICHARDSON | MARCIA WEIDENMIER WATSON



Mc Graw Hill

Introduction to Business Analytics

Introduction to Business Analytics

Vernon J. Richardson
University of Arkansas
Baruch College

Marcia Weidenmier Watson
Trinity University

Mc Graw Hill



INTRODUCTION TO BUSINESS ANALYTICS

Published by McGraw Hill LLC, 1325 Avenue of the Americas, New York, NY 10019. Copyright ©2024 by McGraw Hill LLC. All rights reserved. Printed in the United States of America. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written consent of McGraw Hill LLC, including, but not limited to, in any network or other electronic storage or transmission, or broadcast for distance learning.

Some ancillaries, including electronic and print components, may not be available to customers outside the United States.

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 LWI 28 27 26 25 24 23

ISBN 978-1-266-18914-2
MHID 1-266-18914-9

Cover Image: *Matt Anderson Photography/Getty Images*

All credits appearing on page or at the end of the book are considered to be an extension of the copyright page.

The Internet addresses listed in the text were accurate at the time of publication. The inclusion of a website does not indicate an endorsement by the authors or McGraw Hill LLC, and McGraw Hill LLC does not guarantee the accuracy of the information presented at these sites.

mheducation.com/highered

The McGraw Hill Series in Operations and Decision Sciences

SUPPLY CHAIN MANAGEMENT

Bowersox, Closs, Cooper, and Bowersox
Supply Chain Logistics Management
Sixth Edition

Johnson
Purchasing and Supply Management
Seventeenth Edition

Simchi-Levi, Kaminsky, and Simchi-Levi
Designing and Managing the Supply Chain: Concepts, Strategies, Case Studies
Fourth Edition

Stock and Manrodt
Fundamentals of Supply Chain Management

PROJECT MANAGEMENT

Larson and Gray
Project Management: The Managerial Process
Eighth Edition

SERVICE OPERATIONS MANAGEMENT

Bordoloi, Fitzsimmons, and Fitzsimmons
Service Management: Operations, Strategy, Information Technology
Tenth Edition

MANAGEMENT SCIENCE

Hillier and Hillier
Introduction to Management Science and Business Analytics: A Modeling and Case Studies Approach with Spreadsheets
Seventh Edition

BUSINESS RESEARCH METHODS

Schindler
Business Research Methods
Fourteenth Edition

BUSINESS FORECASTING

Keating and Wilson
Forecasting and Predictive Analytics
Seventh Edition

BUSINESS SYSTEMS DYNAMICS

Sterman
Business Dynamics: Systems Thinking and Modeling for a Complex World

OPERATIONS MANAGEMENT

Cachon and Terwiesch
Operations Management
Third Edition

Cachon and Terwiesch
Matching Supply with Demand: An Introduction to Operations Management
Fifth Edition

Jacobs and Chase
Operations and Supply Chain Management
Seventeenth Edition

Jacobs and Chase
Operations and Supply Chain Management: The Core
Sixth Edition

Schroeder and Goldstein
Operations Management in the Supply Chain: Decisions and Cases
Eighth Edition

Stevenson
Operations Management
Fourteenth Edition

Swink, Melnyk, and Hartley
Managing Operations Across the Supply Chain
Fifth Edition

BUSINESS STATISTICS

Bowerman, Drougas, Duckworth, Froelich, Hummel, Moninger, and Schur
Business Statistics and Analytics in Practice
Ninth Edition

Doane and Seward
Applied Statistics in Business and Economics
Seventh Edition

Doane and Seward
Essential Statistics in Business and Economics
Third Edition

Lind, Marchal, and Wathen
Basic Statistics for Business and Economics
Tenth Edition

Lind, Marchal, and Wathen
Statistical Techniques in Business and Economics
Nineteenth Edition

Jaggia and Kelly
Business Statistics: Communicating with Numbers
Fourth Edition

Jaggia and Kelly
Essentials of Business Statistics: Using Excel
Third Edition

BUSINESS ANALYTICS

Jaggia, Kelly, Lertwachara, and Chen
Business Analytics: Communicating with Numbers
Second Edition

Richardson, Weidenmier Watson
Introduction to Business Analytics

Taddy, Hendrix, Harding
Modern Business Analytics

BUSINESS MATH

Slater and Wittry
Practical Business Math Procedures
Fourteenth Edition

Slater and Wittry
Math for Business and Finance: An Algebraic Approach
Third Edition

Dedications

My most amazing son, Benjamin, who makes me laugh and recognize that life is for experiencing joy. Love you!

—Vern Richardson

My family, whose love and support allow my dreams to come true. Thank you!

—Marcia Weidenmier Watson

About the Authors



Vernon J. Richardson

Vernon J. Richardson is Distinguished Professor of Accounting and the G. William Glezen Chair in the Sam M. Walton College of Business at the University of Arkansas and visiting professor at Baruch College. He received his BS, Master of Accountancy, and MBA from Brigham Young University and his PhD in accounting from the University of Illinois at Urbana–Champaign. He has taught students at the University of Arkansas, University of Illinois, Brigham Young University, and University of Kansas and internationally at the China Europe International Business School (Shanghai), Xi'an Jiaotong Liverpool University, Chinese University of Hong Kong Shenzhen, Aarhus University, and the University of Technology Sydney.

Dr. Richardson is a member of the American Accounting Association and has served as president of the American Accounting Association Information Systems section. He previously served as an editor of *The Accounting Review* and is currently an editor at *Accounting Horizons.* He has published articles in *The Accounting Review, Journal of Information Systems, Journal of Accounting and Economics, Contemporary Accounting Research, MIS Quarterly, International Journal of Accounting Information Systems, Journal of Management Information Systems, Journal of Operations Management,* and *Journal of Marketing.* He is also a co-author of McGraw Hill's *Accounting Information Systems, Introduction to Data Analytics for Accounting,* and *Data Analytics for Accounting* textbooks.



Marcia Weidenmier Watson

Marcia Weidenmier Watson is the Jesse H. Jones Professor Accounting in the Michael Neidorff School of Business at Trinity University. She received her BBA in accounting from the College of William and Mary, and both an MBA in information systems management and accounting and a PhD in accounting from the University of Texas at Austin. She has taught at the University of Texas at Austin, TCU, Mississippi State University, UNC Charlotte, and Trinity University. Prior to receiving her PhD, Dr. Watson was a financial systems consultant for Price Waterhouse, now PricewaterhouseCoopers.

A member of the American Accounting Association, she served as newsletter editor for the American Accounting Association Information Systems section for 16 years. She serves on the editorial boards of *Issues in Accounting Education* and the *Journal of Accounting Education.* She has published articles in *Accounting Horizons, Advances in Management Accounting, Issues in Accounting Education, International Journal of Accounting Information Systems, MIS Quarterly, Journal of Accountancy, Journal of Accounting Education, Journal of Information Systems,* and *Strategic Finance.* She also has had several cases published in *Case Studies in Forensic Accounting and Fraud Auditing.*

From the Authors

Computerization and automation of many business tasks is combining with the explosion of available data to change the way companies work and make decisions. For this reason, business professionals are increasingly required to have an **analytics mindset** to perform their jobs. We recognize that students need to develop the skills to ask the right questions, learn to use common workplace tools (such as Excel®, Tableau®, and Power BI®) to examine and analyze data, and interpret results accurately and effectively to make business decisions.

Developing this analytics mindset early in the study of business is crucial in preparing students to meet the demands of today's workplace. It is also critical in terms of developing business acumen and understanding how basic business functions work. In addition, learning multiple software packages develops technical agility. An analytics mindset, business acumen, and technical agility are essential in preparing students not only for future business classes but also for their internships and post-graduation jobs in the real world.

Introduction to Business Analytics provides a framework for developing a business analytics mindset. This framework, which we call the **SOAR analytics model,** is composed of four steps:

1. Specify the question (Chapter 1)
2. Obtain the data (Chapters 2–3)
3. Analyze the data (Chapters 4, 5, and 11)
4. Report the results (Chapter 6)



This model is used throughout the text in conjunction with the various types of data analysis that analysts need to perform. The lab activities, which appear at the end of each chapter, follow this framework to reinforce the analytical process.

After laying the foundation in Chapters 1–6, we apply the SOAR model in Chapters 7–10 to marketing, accounting, finance, and operations questions. Chapter 11 introduces advanced analytics and discusses how powerful machine learning and other algorithms can improve business analytics. Chapter 12 acts as a capstone, providing three projects that apply the complete SOAR model. The first project asks students to identify the factors that affect Airbnb rental rates, and the second project asks students to analyze LendingClub loans. The third project asks students to use the SOAR framework to address a business question that they have posed.

Vernon J. Richardson
Marcia Weidenmier Watson

Key Features

Focus on Building Skills with Excel®, Tableau,® and Power BI®.

Students learn how to conduct business analytics using Excel, Tableau, and Power BI—three software tools that are widely used by businesses today.

CHAPTER 7 LABS

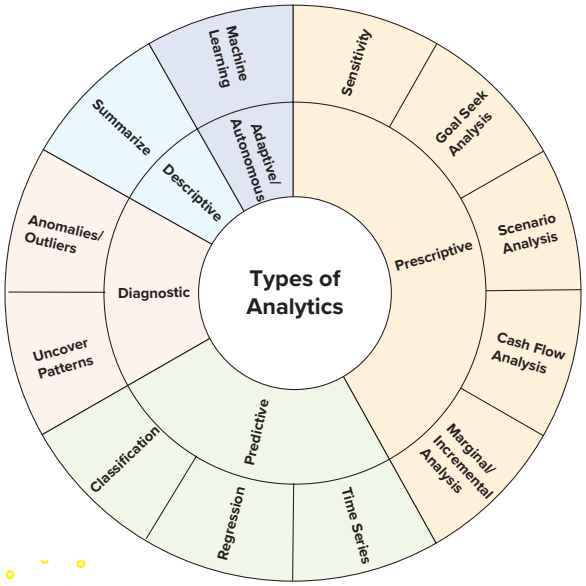
- **LAB 7.1** **Excel:** Descriptive Analytics: Analyzing Company Historical Performance
- **LAB 7.2** **Excel:** Descriptive Analytics: Using a Pivot Table to Analyze Historical Performance by Product Size and Year
- **LAB 7.3** **Tableau:** Descriptive Analytics: Using a Histogram to Evaluate Process Time
- **LAB 7.3** **Power BI:** Descriptive Analytics: Using a Histogram to Evaluate Process Time
- **LAB 7.4** **Excel:** Diagnostic Analytics: Analyzing the Steps in the Sales Process with a Sales Funnel Chart
- **LAB 7.5** **Tableau:** Diagnostic Analytics: Examining Pricing Strategy with Cluster Analysis
- **LAB 7.6** **Excel:** Predictive Analytics: Predicting Sales Revenue from Advertising Expense
- **LAB 7.6** **Tableau:** Predictive Analytics: Predicting Sales Revenue from Advertising Expense
- **LAB 7.7** **Excel:** Prescriptive Analytics: Calculating Internet CPM Rate Using Goal Seek
- **LAB 7.8** **Excel:** Prescriptive Analytics: Calculating Product Price Using Goal Seek

Focus on Building Critical Thinking Skills.

From learning to ask the right questions to interpreting and presenting results, *Introduction to Business Analytics* fosters critical thinking and develops business analysis skills. It teaches students how to use descriptive, diagnostic, predictive, prescriptive, and advanced analytics to answer the following business questions:

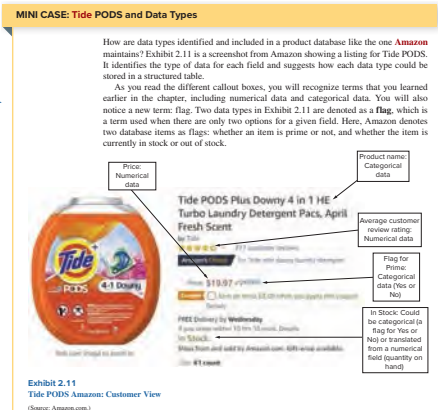
1. **Descriptive analytics:** What happened? What is happening? (Chapter 4)
2. **Diagnostic analytics:** Why did it happen? What are the causes of past results? Why are the results different than expectations? (Chapter 4)
3. **Predictive analytics:** Will it happen in the future? What is the probability something will happen? Can we forecast what will happen? (Chapter 5)
4. **Prescriptive analytics:** What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints? (Chapter 5)
5. **Adaptive/autonomous analytics:** How can we continuously learn using artificial intelligence? Can we learn from past and current events with adaptive learning? (Chapter 11)

The following diagram summarizes the techniques taught.



Emphasis on the Real-World Use of Data.

Mini Cases in each chapter ask students to consider real-world companies and how they can use data to inform their decision-making.



MINI CASE: Tide PODS and Data Types

How are data types identified and included in a product database like the one **Amazon** maintains? Exhibit 2.11 is a screenshot from Amazon showing a listing for Tide PODS. It identifies the type of data for each field and suggests how each data type could be stored in a structured table.

As you read the different callout boxes, you will recognize terms that you learned earlier in the chapter, including numerical data and categorical data. You will also notice a new term: flag. Two data types in Exhibit 2.11 are denoted as a **flag**, which is a term used when there are only two options for a given field. Here, Amazon denotes two database items as flags: whether an item is prime or not, and whether the item is currently in stock or out of stock.

Exhibit 2.11
Tide PODS Amazon: Customer View
(Source: Amazon.com.)

Focus on Data Visualization.

The text emphasizes the creation and interpretation of various types of data visualizations useful in summarizing data and making decisions, including histograms, line graphs, pie charts, and scatterplots.



DATA VISUALIZATION



Lab Exhibit 6.5P.6
Microsoft Power BI

Hands-on Labs.

Introduction to Business Analytics offers more than 60 hands-on labs, each using Excel, Tableau, or Power BI. Each lab has two data sets. The first is used with the step-by-step instructions (with screenshots) presented in the text. The second, alternate data set gives students the opportunity to apply what they learned by using the first data set. Lab assessment appears in Connect through the use of multiple-choice questions. Video tutorials of the Labs are also available in Connect.

Lab 7.2 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Descriptive Analytics: Using a Pivot Table to Analyze Historical Performance by Product Size and Year

Keywords

Descriptive Analytics, Pivot Table

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" Lab 7.1 used the Analysis ToolPak in Excel to generate descriptive analytics for marketing data. This lab uses Excel pivot tables to summarize the data at different levels, allowing drill down and drill up (also known as roll up).

Several years ago, Rob built a table for his son to use in building LEGO creations. The table allowed his son to stand while building, provided a frame to hold LEGO base plates, and got the LEGOs off the floor. Several of Rob's neighbors saw the table and asked him to build tables for their children. The tables grew in popularity and Rob eventually quit his job in 2018 and founded LeTable Inc. He now works full-time building custom LEGO tables.

Customers select from four different sizes of tables:

1. Small, which holds 4 base plates
2. Medium, which holds 8 base plates
3. Large, which holds 16 base plates
4. Deluxe, which has multiple levels as well as conduits for electric lights

Progress Checks.

Progress Check questions posed at key points in each chapter encourage students to consider and apply the concepts presented.

PROGRESS CHECK

6. Which types of data that are internal to the company would be useful in preparing a sales forecast for the next quarter? Which types of data that are external to the company would be meaningful for the same purpose?
7. How can a business analyst use data from the U.S. Census Bureau to understand the demographics of a company's customer base?
8. How can companies improve their traditional business analytics by including Big Data sources such as social media or data from the Internet of Things?

ETHICS

Ethical Use of Data.

Each chapter includes a discussion of important questions related to the ethical collection, use, and sale of data.

Checklist for Creating Effective Charts That Clearly Answer Business Questions

In *How Charts Lie,* Alberto Cairo describes the many ways that a chart might lie, confuse, manipulate, and mislead.[4] Culprits include:

- poor design
- the use of incorrect or an inappropriate amount of data
- the concealment of data
- the suggestion of misleading patterns
- support for pre-existing desired outcomes, opinions, or assumptions
- unclear communication of uncertainty.

For example, the cone of uncertainty used by the U.S. National Hurricane Center (USNC) indicates the likelihood of a hurricane path. The less certain meteorologists are of the hurricane's path, the wider the cone. However, most people incorrectly interpret the map as depicting how the hurricane will strengthen over time, which means that the USNC's chart is problematic.

To become a skilled and ethical creator of data visualizations, Cairo recommends examining many charts and graphs while "being aware of . . . biases and learning to see what charts and graph *don't* necessarily show." [5] He also recommends applying the Golden Rule when you are creating charts: If you don't like being tricked, don't trick other people.

As you think about the data visualizations that you will use to convey information, consider the following questions: How often do you think people deliberately use charts and other data visualizations to mislead or deceive others? What questions should you ask about charts created by others? What questions should you ask about your own charts before you make them public? How might your chart be misinterpreted? Why should you look at the data underlying a chart?

End-of-Chapter Assessment.

The end-of-chapter assignments include real-world application questions, with an emphasis on skills and tools. Each chapter offers discussion questions, exercises, and problems to reinforce learning.

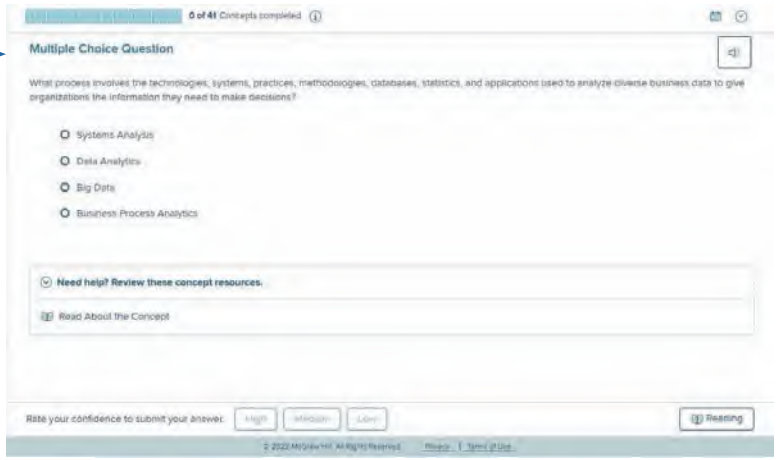
Problems

1. (LO 7.1, 7.4, 7.5, 7.6, 7.7) This chapter focuses on the traditional four Ps of the marketing mix (product, price, place, and promotion) and briefly introduced additional Ps: planning, process, people, and physical evidence. Process is the flow of activities that takes place when the company interacts with a customer. For example, the activities to return a product to **Home Depot** include: (1) Customer enters the Customer Service Desk line; (2) employee listens to customer request; (3) customer presents an item to return; (4) employee verifies the receipt and the item condition; (5) employee approves the return, obtaining manager help/approval when necessary; (6) employee enters information into the system including the reason for the return; and (7) employee provides the customer with a refund. Identify questions related to this process that Home Depot could ask that would require descriptive, diagnostic, predictive, and prescriptive analytics (one question for each type).
2. (LO 7.1, 7.4, 7.5, 7.6, 7.7) This chapter focuses on the traditional four Ps of the marketing mix (product, price, place, and promotion) and briefly introduced additional Ps: planning, process, people, and physical evidence. People refers to having the proper employees with appropriate skills sets or leadership skills. For example, the **Ford Motor Company** needs great salespeople to sell its cars and trucks. Identify questions related to the sales force that Ford could ask that would require descriptive, diagnostic, predictive, and prescriptive analytics to answer (one question for each type).
3. (LO 7.4, 7.5, 7.6, 7.7) **Panera Bread** sends out emails to customers to entice them to come and purchase food. Companies can tell if (and when) customers open emails. Assume that Panera Bread wants to use regression to determine if email open rates are related to subsequent food purchases. What are the independent and dependent variables in the regression? What additional variables would you suggest adding to the regression?
4. (LO 7.5, 7.6) This chapter describes sources of marketing data, including customers' web search history, social media posts, and purchase history. Analyzing this information can reveal private information, such as financial problems, illness, and pregnancy. The following story appeared in both *Fortune* and *The New York Times,* but it has never been confirmed by **Target**. The story provides a good opportunity to understand what could potentially be done with marketing data. A father found out that his teenage daughter was pregnant when Target began sending coupons for baby products to her (based on her purchases).[28] What type of diagnostic and predictive analytics could Target have used to determine that the teenager was pregnant? How can each of these techniques be used in marketing?

Available in Connect

SmartBook®

Is the market-leading adaptive study resource that is proven to strengthen memory recall, increase retention, and boost grades. SmartBook 2.0 identifies and closes knowledge gaps through a continually adapting reading and questioning experience that helps students master the key concepts in each chapter.



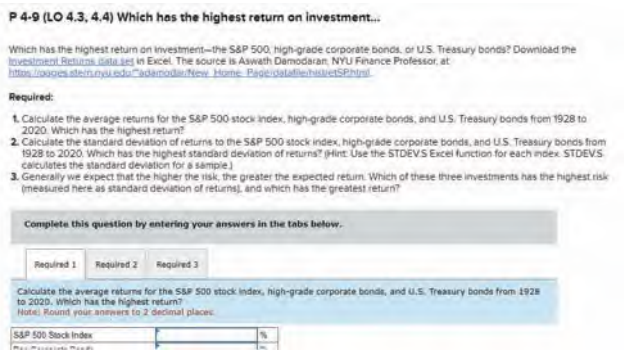
Lecture Videos:

These video-based tutorials are designed to reinforce select chapter concepts.



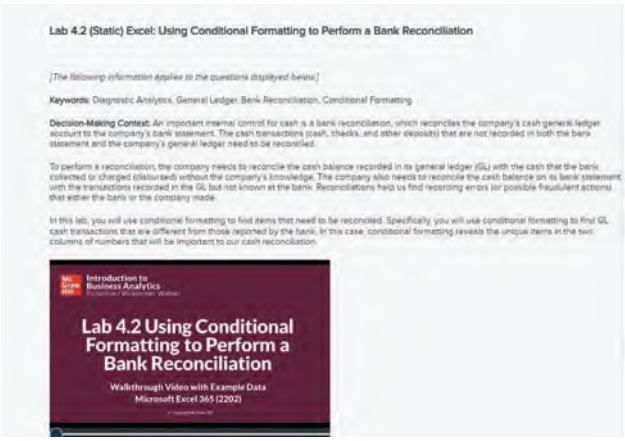
Exercises/Problems:

Select exercises and problems from the text are available for assignment in Connect to ensure students are building an analytical skill set.



Labs with Lab Assessments:

While the labs require students to work outside of Connect in Excel, Tableau, and/or Power BI, Connect allows students to upload their results and answer analytical questions designed to reinforce the lessons from each chapter.



Lab Help Videos:

Help videos for each lab provide a step-by-step tutorial that walks students through the assigned analysis tasks in Excel, Tableau, and Power BI.

Instructors

Student Success Starts with You

Tools to enhance your unique voice

Want to build your own course? No problem. Prefer to use an OLC-aligned, prebuilt course? Easy. Want to make changes throughout the semester? Sure. And you'll save time with Connect's auto-grading, too.

65%
Less Time Grading



Laptop: Getty Images; Woman/dog: George Doyle/Getty Images

A unique path for each student

In Connect, instructors can assign an adaptive reading experience with SmartBook® 2.0. Rooted in advanced learning science principles, SmartBook 2.0 delivers each student a personalized experience, focusing students on their learning gaps, ensuring that the time they spend studying is time well-spent.
mheducation.com/highered/connect/smartbook

Affordable solutions, added value

Make technology work for you with LMS integration for single sign-on access, mobile access to the digital textbook, and reports to quickly show you how each of your students is doing. And with our Inclusive Access program, you can provide all these tools at the lowest available market price to your students. Ask your McGraw Hill representative for more information.

Solutions for your challenges

A product isn't a solution. Real solutions are affordable, reliable, and come with training and ongoing support when you need it and how you want it. Visit **supportateverystep.com** for videos and resources both you and your students can use throughout the term.

SUPPORT AT every step

Students
Get Learning that Fits You

Effective tools for efficient studying

Connect is designed to help you be more productive with simple, flexible, intuitive tools that maximize your study time and meet your individual learning needs. Get learning that works for you with Connect.

Study anytime, anywhere

Download the free ReadAnywhere® app and access your online eBook, SmartBook® 2.0, or Adaptive Learning Assignments when it's convenient, even if you're offline. And since the app automatically syncs with your Connect account, all of your work is available every time you open it. Find out more at **mheducation.com/readanywhere**



iPhone: Getty Images

"I really liked this app—it made it easy to study when you don't have your textbook in front of you."

- Jordan Cunningham, Eastern Washington University



Everything you need in one place

Your Connect course has everything you need—whether reading your digital eBook or completing assignments for class, Connect makes it easy to get your work done.

Learning for everyone

McGraw Hill works directly with Accessibility Services Departments and faculty to meet the learning needs of all students. Please contact your Accessibility Services Office and ask them to email accessibility@mheducation.com, or visit **mheducation.com/about/accessibility** for more information.

Also Available with Connect

Proctorio

Remote Proctoring & Browser-Locking Capabilities

Remote proctoring and browser-locking capabilities, hosted by Proctorio within Connect, provide control of the assessment environment by enabling security options and verifying the identity of the student.

Seamlessly integrated within Connect, these services allow instructors to control the assessment experience by verifying identification, restricting browser activity, and monitoring student actions.

Instant and detailed reporting gives instructors an at-a-glance view of potential academic integrity concerns, thereby avoiding personal bias and supporting evidence-based claims.

ReadAnywhere®

Read or study when it's convenient for you with McGraw Hill's free ReadAnywhere® app. Available for iOS or Android smartphones or tablets, ReadAnywhere gives users access to McGraw Hill tools including the eBook and SmartBook® 2.0 or Adaptive Learning Assignments in Connect. Take notes, highlight, and complete assignments offline—all of your work will sync when you open the app with Wi-Fi access. Log in with your McGraw Hill Connect username and password to start learning—anytime, anywhere!

OLC-Aligned Courses

Implementing High-Quality Instruction and Assessment through Preconfigured Courseware

In consultation with the Online Learning Consortium (OLC) and our certified Faculty Consultants, McGraw Hill has created pre-configured courseware using OLC's quality scorecard to align with best practices in online course delivery. This turnkey courseware contains a combination of formative assessments, summative assessments, homework, and application activities, and can easily be customized to meet an individual instructor's needs and desired course outcomes. For more information, visit https://www.mheducation.com/highered/olc.

Tegrity: Lectures 24/7

Tegrity in Connect is a tool that makes class time available 24/7 by automatically capturing every lecture. With a simple one-click start-and-stop process, you can capture all computer screens and corresponding audio in a format that is easy to search, frame by frame. Students can replay any part of any class with easy-to-use, browser-based viewing on a PC, Mac, or mobile device.

Educators know that the more students can see, hear, and experience class resources, the better they learn. In fact, studies prove it. Tegrity's unique search feature helps students efficiently find what they need, when they need it, across an entire semester of class recordings. Help turn your students' study time into learning moments immediately supported by your lecture. With Tegrity, you also increase intent listening and class participation by easing students' concerns about note-taking. Using Tegrity in Connect will make it more likely you will see students' faces, not the tops of their heads.

Test Builder in Connect

Available within Connect, Test Builder is a cloud-based tool that enables instructors to format tests that can be printed, administered within a Learning Management System, or exported as a Word document. Test Builder offers a modern, streamlined interface for easy content configuration that matches course needs, without requiring a download.

Test Builder allows you to:

- access all test bank content from a particular title.
- easily pinpoint the most relevant content through robust filtering options.
- manipulate the order of questions or scramble questions and/or answers.
- pin questions to a specific location within a test.
- determine your preferred treatment of algorithmic questions.
- choose the layout and spacing.
- add instructions and configure default settings.

Test Builder provides a secure interface for better protection of content and allows for just-in-time updates to flow directly into assessments.

Writing Assignment

Available within Connect and Connect Master, the Writing Assignment tool delivers a learning experience to help students improve their written communication skills and conceptual understanding. As an instructor, you can assign, monitor, grade, and provide feedback on writing more efficiently and effectively.

Create



Your Book, Your Way

McGraw Hill's Content Collections Powered by Create® is a self-service website that enables instructors to create custom course materials—print and eBooks—by drawing upon McGraw Hill's comprehensive, cross-disciplinary content. Choose what you want from our high-quality textbooks, articles, and cases. Combine it with your own content quickly and easily, and tap into other rights-secured, third-party content such as readings, cases, and articles. Content can be arranged in a way that makes the most sense for your course, and you can include the course name and information as well. Choose the best format for your course: color print, black-and-white print, or eBook. The eBook can be included in your Connect course and is available on the free ReadAnywhere® app for smartphone or tablet access as well. When you are finished customizing, you will receive a free digital copy to review in just minutes! Visit McGraw Hill Create®—www.mcgrawhillcreate.com—today and begin building!

Reflecting the Diverse World Around Us

McGraw Hill believes in unlocking the potential of every learner at every stage of life. To accomplish that, we are dedicated to creating products that reflect, and are accessible to, all the diverse, global customers we serve. Within McGraw Hill, we foster a culture of belonging, and we work with partners who share our commitment to equity, inclusion, and diversity in all forms. In McGraw Hill Higher Education, this includes, but is not limited to, the following:

- Refreshing and implementing inclusive content guidelines around topics including generalizations and stereotypes, gender, abilities/disabilities, race/ethnicity, sexual orientation, diversity of names, and age.
- Enhancing best practices in assessment creation to eliminate cultural, cognitive, and affective bias.
- Maintaining and continually updating a robust photo library of diverse images that reflect our student populations.
- Including more diverse voices in the development and review of our content.
- Strengthening art guidelines to improve accessibility by ensuring meaningful text and images are distinguishable and perceivable by users with limited color vision and moderately low vision.

Acknowledgments

Our sincere thanks to all who helped us on this project. We also thank Katie Terrell for helping shape our vision for this book.

Our biggest thanks to the awesome team at McGraw Hill, including Rebecca Olson, Eric Weber, Steven Rigolosi, Katherine Ward, Harper Christopher, Kelsey Darin, and Jill Eccher.

We also thank Shane Lunceford from LendingClub for help us gain access to LendingClub data. We also acknowledge Bonnie Klamm, Mike Casey, and Edie Schmidt for help with accuracy checking.

In addition, we appreciate the contributions of the following reviewers, who provided ideas and insights for this new book.

Shikhar Acharya
Loras College

Guclu Atinc
Texas A& M University–Commerce

John H. Batchelor
University of West Florida

Palash Bera
Saint Louis University

Carol Cagle
Mercer University

Cenk Caliskan
Utah Valley University

Roberto Cavazos
University of Baltimore

Xiaohui Chang
Oregon State University

Jehanzeb Cheema
George Mason University

Julie Dilling
Moraine Park Technical College

Christie Fuller
Boise State University

Regina Gramling
Auburn University

Daryl D. Green
Oklahoma Baptist University

Bo Han
Texas A&M University–Commerce

Keith Harman
Oklahoma Baptist University

Faria Huq
Lake Erie College

Jaehwan Jeong
Radford University

Patrick Johanns
University of Iowa

Lauren Kleitz
Xavier University

Brian Kovar
Kansas State University

Jooh Lee
Rowan University

Anita Lee-Post
University of Kentucky

Vic Matta
Ohio University

Jack McCann
Purdue University

Miriam McGaugh
Oklahoma State University

Reza Mousavi
Western Michigan University

Anita Mukherjee
University of Wisconsin–Madison

Hongsong Neuhauser
Elizabethtown College

Kristen Pettey
Southwestern College

Richard R. J. Podeschi
Millikin University

Roman Rabinovich
Boston University

Carole Shook
University of Arkansas

Mike Shurden
Lander University

B. S. Vijayaraman
The University of Akron

Benjamin Williams
University of Denver

Emre Yetgin
Rider University

Amy J. N. Yurko
Duquesne University

Faye X. Zhu
Rowan University

Brief Table of Contents

About the Authors vii

From the Authors viii

Key Features x

Available in Connect xiii

Acknowledgments xix

Chapter 1 Specify the Question: Using Business Analytics to Address Business Questions 2

Chapter 2 Obtain the Data: An Introduction to Business Data Sources 68

Chapter 3 Analyze the Data: Basic Statistics and Tools Required in Business Analytics 134

Chapter 4 Analyze the Data: Exploratory Business Analytics (Descriptive and Diagnostic Analytics) 220

Chapter 5 Analyze the Data: Confirmatory Business Analytics (Predictive Analytics and Prescriptive Analytics) 272

Chapter 6 Report the Results: Using Data Visualization 336

Chapter 7 Marketing Analytics 438

Chapter 8 Accounting Analytics 540

Chapter 9 Financial Analytics 618

Chapter 10 Operations Analytics 686

Chapter 11 Advanced Business Analytics 774

Chapter 12 Using the SOAR Analytics Model to Put It All Together: Three Capstone Projects 844

Appendix A Excel Tutorial (Formatting, Sorting, Filtering, and Pivot Tables) 860

Appendix B Tableau Tutorial 868

Appendix C Power BI Desktop Tutorial 872

Appendix D Basic Statistics Tutorial 879

Appendix E Installing Excel's Analysis ToolPak Add-In 883

Appendix F Installing Excel's Solver Add-In 885

***GLOSSARY G1**

INDEX I1

*Indicates Connect only material

Table of Contents

Chapter 1

Specify the Question: Using Business Analytics to Address Business Questions 2

1.1 Businesses Create Value 3
The Increasing Availability of Data and the Role of Business Analyst 4
1.2 The Difference Between Data and Information 6
1.3 Summarizing the Role of the Business Analyst 7
1.4 Business Analytics Across the Different Business Functions 8
1.5 An Introduction to the SOAR Analytics Model 9
Illustration of the SOAR Analytics Model: Predicting Inventory Needs 11
S: Specify the Question 11
O: Obtain the Data 12
A: Analyze the Data 14
R: Report the Results 16
1.6 Using Exploratory and Explanatory Data Visualizations in Business Analytics 17
Exploratory Visualizations 17
Explanatory Visualizations 18
Chapter Summary 21
Key Terms 21
Answers to Progress Checks 23
Multiple-Choice Questions 24
Discussion Questions 26
Problems 29
Lab 1.1 Excel: Assessing Profitability of a Mowing Business 32
Lab 1.2 Excel: Assessing Customer Profitability 39
Lab 1.2 Tableau: Assessing Customer Profitability 44
Lab 1.2 Power BI: Assessing Customer Profitability 49
Lab 1.3 Excel: Assessing Product (SKU) Profitability 52
Lab 1.3 Tableau: Assessing Product (SKU) Profitability 58
Lab 1.3 Power BI: Assessing Product (SKU) Profitability 64

Chapter 2

Obtain the Data: An Introduction to Business Data Sources 68

2.1 Internal and External Data Sources 70
Internal Data Sources 70
External Data Sources 72
2.2 Obtaining the Data 77
Text Data 77
Tabular Data 77
2.3 Structured Data Types: Categorical versus Numerical 78
Categorical Data: Nominal and Ordinal Data 79
Numerical Data: Interval and Ratio Data 81
Additional Ways to Classify Data 83
2.4 Preparing Data for Analysis 84
Step 1: Ensure Data Quality 84
Step 2: Validate the Data for Completeness and Integrity 85
Step 3: Cleanse the Data 86
Step 4: Perform Preliminary Exploratory Analysis 89
2.5 Tools Used to Prepare Data for Analysis 90
2.6 Gathering and Protecting Data Ethically 91
Chapter Summary 92
Key Terms 93
Answers to Progress Checks 95
Multiple-Choice Questions 96
Discussion Questions 98
Problems 100
Lab 2.1 Excel: Identifying and Working with Different Data Types 104
Lab 2.2 Tableau: Preparing Different Data Types for Analysis 109
Lab 2.2 Power BI: Preparing Different Data Types for Analysis 113
Lab 2.3 Tableau: Conducting Preliminary Exploratory Analysis 117
Lab 2.3 Power BI: Conducting Preliminary Exploratory Analysis 122
Lab 2.4 Excel: Aggregating and Visualizing Different Data Types 127

Chapter 3

Analyze the Data: Basic Statistics and Tools Required in Business Analytics 134

3.1 Defining Populations and Samples 136
Parameters and Statistics 136
Descriptive/Summary Statistics, Inferential Statistics, and Hypotheses 137
3.2 Sampling Methods, Data Reduction, and Bias 137
Sampling Methods 137

Data Reduction 138
The Decision Process of Using a Subset of the Data 139
Bias in Business Analytics 140
3.3 Understanding Basic Statistics 141
Probability Distributions 141
Measures of Central Tendency 142
Measures of Dispersion 144
Continuous Probability Distributions 146
3.4 Using Software Tools to Create Summary Statistics 148
Excel Functions 148
Excel Data Analysis ToolPak—Descriptive Statistics 149
Changing the Default Aggregation in Excel Pivot Table, Power BI, and Tableau 150
3.5 Interpreting and Visualizing Statistics 153
Interpreting Statistics Through Visualizations 154
3.6 Hypothesis Testing 156
Confidence Intervals: Estimating Population Parameters from Sample Statistics 157
Hypothesis Testing 157
3.7 *t*-Tests, Anova Tests, and Chi-Square Tests 164
t-Tests: Testing for a Difference of Means Between Two Paired (Dependent) Groups 164
ANOVA: Testing for Difference of Means Among Three or More Groups 165
Chi-Square Tests: Testing for a Relationship Between Two or More Variables 166
3.8 Introduction to Correlation and Regression Analysis 167
Correlation 168
Regression Analysis and Line of Best Fit 170
Chapter Supplement 174
Discrete Data Distributions 174
Summary 174
Key Terms 175
Answers to Progress Checks 178
Multiple-Choice Questions 180
Discussion Questions 181
Problems 184
Lab 3.1 Excel: Using Excel Functions to Calculate Descriptive Statistics to Gain Insights About the Distribution of a Sales Data Set 189
Lab 3.2 Excel: Using the Excel Analysis ToolPak to Calculate Descriptive Statistics to Gain Insights About the Distribution of a Sales Data Set 193
Lab 3.2 Tableau: Calculating Descriptive Statistics to Gain Insights About the Distribution of a Sales Data Set 197
Lab 3.3 Excel: Performing a *t*-test for Difference in Means to Determine If the Differences Between In-Person and Online Sales Are Statistically Significant 201
Lab 3.4 Excel: Performing an ANOVA Test for Difference in Means to Determine If There Are Significant Differences Between the Average 4-Year Degree Completion Rate/SAT Average for Public, Private, and For-Profit Colleges 206
Lab 3.5 Excel: Deriving Cost Drivers for Activity-Based Costing (Regression Analysis) 211

Chapter 4

Analyze the Data: Exploratory Business Analytics (Descriptive and Diagnostic Analytics) 220

4.1 The Third Step of the SOAR Analytics Model: Analyze the Data 222
4.2 Matching the Analytics Type to the Business Question Asked 222
4.3 Defining Exploratory and Confirmatory Business Analytics 225
4.4 Descriptive Analytics 226
Systems that Provide Data for Descriptive Analytics 226
Statistical and Summarization Techniques for Descriptive Analytics 228
Using Descriptive Statistics as a Descriptive Analytics Technique 228
Data Visualization in Descriptive Analytics: Graphs and Histograms 229
Examples of Descriptive Analytics: Horizontal and Vertical Analysis of Performance 230
4.5 Going from Descriptive Analytics to Diagnostic Analytics 232
Diagnostic Analytics: Identifying Anomalies and Outliers 232
Finding Previously Unknown Linkages, Patterns, or Relationships Between Variables 238
4.6 Introduction to Techniques Used in Data Analysis 240
The Excel Data Analysis ToolPak 241
Chapter Summary 242
Key Terms 243
Answers to Progress Checks 244
Multiple-Choice Questions 245
Discussion Questions 247
Problems 249
Lab 4.1 Excel: Evaluating Inventory Using Inventory Turnover, Waste, and Profit Margins 253

Lab 4.2 Excel: Using Conditional Formatting to Perform a Bank Reconciliation 258
Lab 4.3 Excel: Applying Benford's Law 262

Chapter 5

Analyze the Data: Confirmatory Business Analytics (Predictive Analytics and Prescriptive Analytics) 272

5.1 Confirmatory Data Analytics 274
5.2 Predictive Analytics 274
Classification 274
Regression 277
Time Series Analysis 279
5.3 Base Rates and Base Rate Fallacy 280
5.4 Prescriptive Analytics 281
Prescriptive Analytics Techniques 281
5.5 Summary of Analytics Performed to Address Business Questions 284
Summary 286
Key Terms 286
Answers to Progress Checks 287
Multiple-Choice Questions 287
Discussion Questions 289
Problems 290
Lab 5.1 Excel: Evaluating the Relationship Between Sales and R&D Expenditures 293
Lab 5.2 Excel: Evaluating the Relationship Between Sales and R&D Expenditures: Testing for a Nonlinear Relationship 299
Lab 5.3 Excel: Forecasting Product Demand Using Time Series Analysis 305
Lab 5.4 Tableau: Forecasting Product Demand Using Time Series Analysis 308
Lab 5.5 Power BI: Forecasting Product Demand Using Time Series Analysis 316
Lab 5.6 Excel: Forecasting Product Demand Using Regression 319
Lab 5.7 Excel: Assessing the Returns to Education: To MBA or Not to MBA? 324
Lab 5.8 Excel: Applying Scenario Analysis: Possible Trade War 331

Chapter 6

Report the Results: Using Data Visualization 336

6.1 The Basics of Data Visualization 338
Why Use Visualizations? 338
Sorting Considerations 340
6.2 Distinguishing Among Chart Types 342
Charts Appropriate for Categorical Data 343
Charts Appropriate for Numerical (Quantitative) Data 347
6.3 Visualizing Exploratory Business Analytics 348
Bar Charts, Line Charts, and Pie Charts 349
Bar Charts Versus Histograms 351
Visualizing Outliers and Anomalies 354
Pivot Tables 355
6.4 Visualizing Confirmatory Analytics 356
Correlation and Regression 357
Forecasting with Time Series Data 358
6.5 Presenting Data in a Dashboard 359
Checklist for Creating Effective Charts That Clearly Answer Business Questions 361
6.6 Communicating Your Data with Words: Executive Summaries and Reports 362
Executive Summary 362
Full Report 363
Chapter Summary 365
Key Terms 365
Answers to Progress Checks 366
Multiple-Choice Questions 367
Discussion Questions 370
Problems 371
Lab 6.1 Excel: Descriptive Analytics: Visualizing Pivot Table Data Using Conditional Formatting and Sparklines 377
Lab 6.2 Tableau: Descriptive Analytics: Analyzing Sales Revenue by Product with a Tree Map 384
Lab 6.2 Power BI: Descriptive Analytics: Analyzing Sales Revenue by Product with a Tree Map 391
Lab 6.3 Tableau: Descriptive Analytics: Analyzing Sales Revenue by Customer with a Bar Chart and Filters 397
Lab 6.3 Power BI: Descriptive Analytics: Analyzing Sales Revenue by Customer with a Bar Chart and Filters 405
Lab 6.4 Excel: Descriptive Analytics: Creating a Dashboard Using Pivot Tables and Slicers 413
Lab 6.5 Tableau: Descriptive Analytics: Creating a Dashboard 421
Lab 6.5 Power BI: Descriptive Analytics: Creating a Dashboard 430

Chapter 7

Marketing Analytics 438

7.1 What Is Marketing? 440
The Marketing Mix 440
7.2 Specifying The Marketing Question 441

7.3 Obtain The Data: What Marketing Data Are Available? A Discussion of Marketing Data Sources 443
Internal Marketing Data 443
External Marketing Data 446
Combining Internal and External Data 449
7.4 Descriptive Marketing Analytics 451
Statistical, Summarization, and Data Visualization Techniques for Descriptive Analytics 451
Examples of Marketing Descriptive Analytics 451
7.5 Diagnostic Marketing Analytics 456
Identification of Anomalies and Outliers 456
Cluster Analysis 458
Correlation and Summarization 459
7.6 Predictive Marketing Analytics 462
Decision Tree 463
7.7 Prescriptive Marketing Analytics 464
Goal-Seek Analysis 465
7.8 Report The Results 468
Chapter Summary 469
Key Terms 470
Answers to Progress Checks 471
Multiple-Choice Questions 473
Discussion Questions 475
Problems 478
Lab 7.1 Excel: Descriptive Analytics: Analyzing Company Historical Performance 482
Lab 7.2 Excel: Descriptive Analytics: Using a Pivot Table to Analyze Historical Performance by Product Size and Year 490
Lab 7.3 Tableau: Descriptive Analytics: Using a Histogram to Evaluate Process Time 497
Lab 7.3 Power BI: Descriptive Analytics: Using a Histogram to Evaluate Process Time 502
Lab 7.4 Excel Diagnostic Analytics: Analyzing the Steps in the Sales Process with a Sales Funnel Chart 507
Lab 7.5 Tableau: Diagnostic Analytics: Examining Pricing Strategy with Cluster Analysis 514
Lab 7.6 Excel: Predictive Analytics: Predicting Sales Revenue from Advertising Expense 520
Lab 7.6 Tableau: Predictive Analytics: Predicting Sales Revenue from Advertising Expense 525
Lab 7.7 Excel: Prescriptive Analytics: Calculating Internet CPM Rate Using Goal Seek 530
Lab 7.8 Excel: Prescriptive Analytics: Calculating Product Price Using Goal Seek 534

Chapter 8

Accounting Analytics 540

8.1 The Role of Accounting in Business 542
The Four Primary Branches of Accounting 542
8.2 Specifying the Accounting Question 543
8.3 Obtain the Data: The Sources of Accounting Data 545
Financial Accounting Data Sources 545
Managerial Accounting Data Sources 548
Auditing Data Sources 551
Tax Data Sources 551
8.4 Descriptive Accounting Analytics 552
8.5 Diagnostic Accounting Analytics 556
Identifying Anomalies and Outliers 556
Performing Drill-Down Analysis to Determine Relations, Patterns, and Linkages Among Variables 559
8.6 Predictive Accounting Analytics 561
Using Time Series Analysis to Predict Sales 561
8.7 Prescriptive Accounting Analytics 562
Goal-Seek Analysis 563
Sensitivity Analysis 564
8.8 Report The Results 565
Tables Are Sometimes a More Effective Way to Show Results 565
Using Visualizations to Highlight Anomalies 566
Using a Graph to Show Break-Even Sales 567
Summary 567
Key Terms 568
Answers to Progress Checks 569
Multiple-Choice Questions 570
Discussion Questions 571
Problems 574
Lab 8.1 Excel: Descriptive Analytics: Performing Horizontal Analysis 577
Lab 8.2 Excel: Descriptive Analytics: Performing Vertical Analysis 582
Lab 8.3 Excel: Diagnostic Analytics: Using Fuzzy Matching to Look for Fraud 586
Lab 8.4 Excel: Diagnostic Analytics: Estimating Fixed and Variable Costs 592
Lab 8.5 Excel: Predictive Analytics: Forecasting Future Performance of IBM 598
Lab 8.6 Tableau: Predictive Analytics: Forecasting Future Performance of IBM 602
Lab 8.7 Power BI: Predictive Analytics: Forecasting Future Performance of IBM 611

Lab 8.8 Excel: Prescriptive Analytics: Using Goal-Seek Analysis to Determine the Break-Even Point 614

Chapter 9

Financial Analytics 618

9.1 The Role of Finance in Business 620
The Three Primary Branches of Finance 620
9.2 Specifying the Finance Question 621
9.3 Obtain the Data: Financial Data Sources 622
Stock-Return Data 622
Summarized Financial Data 623
Financial Statement Data 623
9.4 Descriptive Financial Analytics 625
Statistical and Summarization Techniques for Descriptive Analytics 625
Descriptive Statistics 626
9.5 Diagnostic Financial Analytics 627
Finding Anomalies and Outliers 627
Drill-Down Analytics Using DuPont Ratios 627
Abbreviated DuPont Ratio Analysis for Return on Assets 629
Risk/Return and the Sharpe Ratio 631
Risk/Return Diagnostic Analysis Using the Sharpe Ratio 632
Risk/Return Diagnostic Analysis Using Regression Analysis 632
9.6 Predictive Financial Analytics 633
Altman's Z and Bankruptcy Classification 634
Predicting Who Will Be Offered a Loan 635
9.7 Prescriptive Financial Analytics 638
Prescriptive Cash Flow Analysis 639
Sensitivity Analysis 644
9.8 Report The Results 645
Line Graphs to Compare Performance 645
Research Reports and Earnings Forecasts 645
Line Graphs in Sensitivity Analysis 647
Summary 647
Key Terms 648
Answers to Progress Checks 649
Multiple-Choice Questions 650
Discussion Questions 652
Problems 654
Lab 9.1 Excel: Descriptive and Diagnostic Analytics: Calculating Returns to Investments Using the Sharpe Ratio 659
Lab 9.2 Excel: Diagnostic Analytics: Applying DuPont Analysis of Financial Performance 664
Lab 9.3 Excel: Predictive Analytics: Evaluating Loan Acceptance 667
Lab 9.4 Excel: Predictive Analytics: Predicting Bankruptcy 671
Lab 9.5 Excel: Prescriptive Analytics: Evaluating Investments Using NPV 678
Lab 9.6 Excel: Prescriptive Analytics: Evaluating Investments Using IRR 682

Chapter 10

Operations Analytics 686

10.1 The Role of Operations in Business 688
The Three Primary Branches of Operations 688
10.2 Specifying the Operations Question 690
10.3 Obtain the Data: Operations Data Sources 692
Human Resources Data 692
IT Operations Data 694
Supply Chain Data 695
10.4 Descriptive Operations Analytics 697
Statistical, Summarization, and Data Visualization Techniques 697
10.5 Diagnostic Operations Analytics 702
10.6 Predictive Operations Analytics 707
10.7 Prescriptive Operations Analytics 710
Optimization 711
10.8 Report The Results 714
Chapter Summary 716
Key Terms 716
Answers to Progress Checks 717
Multiple-Choice Questions 719
Discussion Questions 721
Problems 723
Lab 10.1 Excel: Descriptive Analytics/Supply Chain: Calculating KPIs for Walmart and Amazon 728
Lab 10.2 Tableau: Descriptive Analytics/Human Resources: Analyzing Employee Turnover Using a Bar Chart 734
Lab 10.2 Power BI: Descriptive Analytics/ Human Resources: Analyzing Employee Turnover Using a Bar Chart 739
Lab 10.3 Excel: Diagnostic Analytics/Supply Chain: Analyzing KPI Differences Between Walmart and Amazon 745
Lab 10.4 Tableau: Diagnostic Analytics/IT Operations: Analyzing Website Response Time Using a Control Chart 750

Lab 10.5 Tableau: Predictive Analytics/Supply Chain: Forecasting Unit Demand 758
Lab 10.6 Excel: Prescriptive Analytics/ Human Resources: Optimizing Employee Assignment 768

Chapter 11

Advanced Business Analytics 774

11.1 Adaptive/Autonomous Business Analytics 776
Adaptive/Autonomous Analytics: Mimicking Human Capabilities 777
11.2 Advanced Analytic Methods 780
Artificial Intelligence 780
11.3 Adaptive/Autonomous Analytics in the Business Functions 783
Marketing Analytics 784
Accounting Analytics 786
Financial Analytics 789
Operations Analytics 791
11.4 Trending Technologies 798
Robotic Process Automation (RPA) 799
Cockpits 800
Voice and Image Analytics 800
Blockchain Analytics 801
11.5 Data-Driven Organizations and the Future of Business Analytics 805
Limitations of Business Analytics 806
The Future of Business Analytics 806
Summary 807
Key Terms 808
Answers to Progress Checks 809
Multiple-Choice Questions 810
Discussion Questions 812
Problems 814
Lab 11.1 Excel: Advanced Analytics/Marketing: Analyzing Customer Repurchase Rates Using Cohorts 821
Lab 11.2 Tableau: Advanced Analytics/ Marketing: Analyzing Text Using a Word Cloud 828
Lab 11.3 Excel: Advanced Analytics/ Human Resources: Visualizing Survival Analysis 831
Lab 11.4 Excel: Advanced Analytics/Supply Chain: Estimating Demand Using Monte Carlo Simulation 836
Lab 11.5: Advanced Analytics: Understanding Blockchain 840

Chapter 12

Using the SOAR Analytics Model to Put It All Together

Three Capstone Projects 844

12.1 Project 1: Determining the Price of Airbnb Nightly Rentals in New York City 846
Specify the Question 847
Obtain the Data 847
Analyze the Data 850
Report the Results 851
12.2 Project 2: Determining the Factors Associated With Loan Repayment 851
Specify the Question 851
Obtain the Data 851
Analyze the Data 853
Report the Results 855
12.3 Project 3: Completing your own Project Using the SOAR Analytics Model 855
Using the SOAR Analytics Model 855
Deliverables 856

Appendix A

Excel Tutorial (Formatting, Sorting, Filtering, and Pivot Tables) 860

Appendix B

Tableau Tutorial 868

Appendix C

Power BI Desktop Tutorial 872

Appendix D

Basic Statistics Tutorial 879

Appendix E

Installing Excel's Analysis ToolPak Add-In 883

Appendix F

Installing Excel's Solver Add-In 885

***GLOSSARY Online G1**

INDEX I1

*Indicates Connect only material

Introduction to Business Analytics

Chapter 1

Specify the Question:
Using Business Analytics to Address Business Questions

A Look at This Chapter

- Businesses, which exist to create value, collect and have access to enormous amounts of data related to their business processes.
- Business analysts use data to help managers answer questions about value creation and other business functions.
- The SOAR analytics model is a useful framework for performing business analytics and advising managers.
- The first step of the SOAR model is SPECIFY THE QUESTION.

A Look Ahead

- To address business questions, we must know what data are available, how to access those data, and how to evaluate those data.
- Chapter 2 introduces the second step in the SOAR model: OBTAIN THE DATA. It summarizes the sources of business data, the various types of data, and key data management tools.



Joe Marino-Bill Cantrell/UPI/Alamy Stock Photo

Discovering the Value in Data

Tesla recently announced it had made significant mistakes in calculating the true cost of its Solar Roof, which is composed of solar-powered roofing tiles for houses. The Solar Roof tiles are more appealing to the eye than traditional solar panels. The result was short-term losses for Tesla and ultimately significant price hikes for Tesla customers.[1]

Business analysts use data to address management's questions. Perhaps Elon Musk, the CEO of Tesla, should have asked different or better questions to determine the correct cost for its solar roofs to help determine the pricing that would have generated profits instead of losses. Could business analysts at Tesla have performed more or better data analysis to avoid the company's expensive mistake?

In this chapter, we define the role of business analysts, who analyze data to address management questions for use in decision-making.

[1] *Business Insider*, https://www.businessinsider.com/elon-musk-tesla-solar-roof-price-mistakes-calculating-cost-2021.4, accessed 04/30/2021.

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

- **LO 1.1:** Define a business process and explain why increased data availability has given rise to the role of the business analyst.
- **LO 1.2:** Differentiate between data and information.
- **LO 1.3:** Summarize the role of the business analyst.
- **LO 1.4:** Describe how the various business functions use business analytics.
- **LO 1.5:** Identify the components of the SOAR analytics model.
- **LO 1.6:** Describe the use of exploratory data visualizations and explanatory data visualizations.

LABS FOR CHAPTER 1

- **LAB 1.1 Excel:** Assessing Profitability of a Mowing Business
- **LAB 1.2 Excel:** Assessing Customer Profitability
- **LAB 1.2 Tableau:** Assessing Customer Profitability
- **LAB 1.2 Power BI:** Assessing Customer Profitability
- **LAB 1.3 Excel:** Assessing Product (SKU) Profitability
- **LAB 1.3 Tableau:** Assessing Product (SKU) Profitability
- **LAB 1.3 Power BI:** Assessing Product (SKU) Profitability

LO 1.1

Define a business process and explain why increased data availability has given rise to the role of the business analyst.

1.1 BUSINESSES CREATE VALUE

A company earns revenue by using inputs such as raw materials, talented workers, buildings, and equipment to produce and then sell a more valuable output, such as completed architectural plans or laptop computers. Consider **Apple** as an example. By designing, manufacturing, and selling iPhones to the general public, Apple creates value because the iPhone is a powerful computing and communications device. If Apple did not create value, ultimately it would not survive.

Also consider some other examples. Manufacturing companies such as **Tesla** create value by procuring and processing raw materials to produce more valuable products, including electric



Sundry Photography/Shutterstock

cars, for their customers. Retailers such as **Amazon** create value by procuring products and selling them in a way that's convenient for customers, either at a physical location or on an e-commerce platform (website). Amazon enhances the customer experience by providing information about these products through customer reviews and by facilitating convenient payment processes and delivery possibilities. **Bank of America** uses technology and access to capital to create value by offering mortgages and other financial services to its customers.

As you can see, businesses provide value to their customers in many ways. **Business value** refers to all the items, events, and interactions that determine a company's financial health. This value may come from the company's suppliers, customers, employees, and/or information systems. A common measure of a business's value is its increase in long-term profitability. This metric incorporates not only the revenues associated with a product or service but also the cost of producing that product or service because, to stay in business, a firm must earn more in revenues than it incurs in expenses.

To consider how value is created, we begin by looking at the specific activities, or business processes, that a company performs. A **business process** is a coordinated, standardized set of activities conducted by both people and equipment to accomplish a specific business task. Consider the following examples of business processes at various companies.

- **Amazon** accepts an order for a *New York Times* bestseller.
- **KPMG** offers tax advice to client.
- **TruGreen** bills a customer for herbicide sprayed on that customer's lawn.
- **Tesla** produces an electric vehicle to sell in China.
- **Fulton Homes** writes a contract with a customer to build a house with materials printed by a 3D printer.
- **eBay** advertises a hard-to-find classic toy to potential buyers.
- **Toyota** procures parts that will be used in producing its Camry model.
- **DoorDash** pays its drivers, who are independent contractors, for making deliveries.
- **Procter & Gamble (P&G)** addresses customer complaints, sometimes by giving refunds.
- **Boston Consulting Group** summarizes various management compensation packages for a client.

Companies perform thousands, and sometimes millions, of processes each day. Management's challenge is to identify the processes that create the most value and minimize the costs of those processes. To help them with decision-making, managers increasingly seek the assistance of business analysts, who use data to answer managers' questions. A **business analyst** is a data specialist who curates and uses data to help an organization make effective business decisions.

The Increasing Availability of Data and the Role of Business Analyst

The amount of data we generate as a society is enormous. About 2.5 quintillion bytes of data are created each day, and the rate of data growth continues to accelerate. Exhibit 1.1 summarizes this surge of data, which is projected to expand from about 1 zettabyte (ZB) in 2010 to approximately 175 ZB by 2025. (One zettabyte equals one sextillion, or 1,000,000,000,000,000,000,000, bytes of data.) At this rate, how much data we will have by 2050?

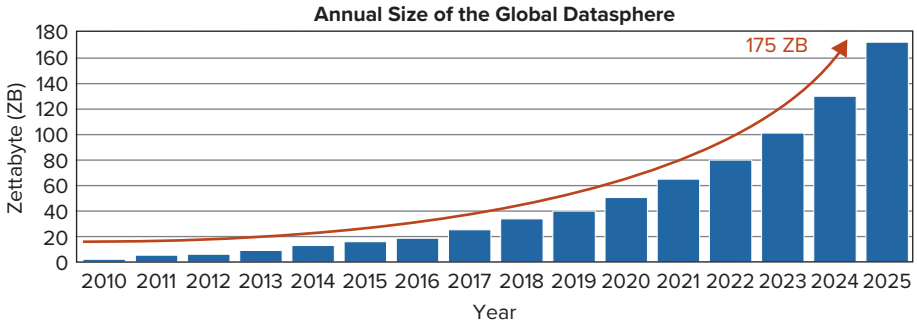


Exhibit 1.1
Data Growth Since 2010 (Actual and Projected)

(Source: IDC Global DataSphere, Nov. 2018, page 6, https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-data-age-whitepaper.pdf, accessed 12/30/2020)

The enormous amount of available data can both help and hinder the work of business analysts. On the positive side, the abundance of data can be invaluable in helping them address managers' questions and offer insight into the company's problems and challenges. On the negative side, the vast amount of data may result in **data overload**, which can prevent business analysts from properly synthesizing and interpreting the data.

With so much data available and computers doing much of the basic collection and simple data analysis, it is increasingly critical that businesses train their employees to develop an **analytics mindset**. This mindset entails the willingness and ability to specify which business questions need to be addressed, find and extract pertinent data that might address those questions, analyze those data, and then report the results to decision-makers. Business analysts are uniquely positioned to perform this type of analysis because they typically understand (1) the questions that their business and its decision-makers are asking and (2) the nature and quality of the business's data. Indeed, according to the consulting firm **EY**, "Analytics is at the heart of every business decision."[2] In this text, you will work to develop an analytics mindset and prepare yourself to take on the new and evolving role of business analyst.

Data Analytics and Your Career

Business analysis is a vibrant career opportunity. According to the U.S. Bureau of Labor Statistics (BLS), 907,600 people were employed as management or business analysts in 2020, with this number expected to grow 14 percent annually from 2020 to 2030. This rate is much faster than the average for all occupations in the United States. In May 2020, the median income for all management analysts, including business analysts, was $87,660.[3]

PROGRESS CHECK

1. How does increasing the amount of data available to address business questions both help and hinder the business analyst's role?
2. Why are business analysts uniquely positioned to address an organization's business questions?

[2]Ernst & Young Foundation: E&Y Academic Resource Center (EYARC). 2017. The Analytics Mindset. Available online on November 17, 2018, at http://aaahq.org/Education/Webinars/6-7-17-EY-Academic-Resource-Center-An-Overview-of-Analytics-Mindset-Competencies-and-Case-Offerings.

[3]U.S. Bureau of Labor Statistics, May 2020, https://www.bls.gov/ooh/business-and-financial/management-analysts.htm. Accessed September 2021.

LO 1.2

Differentiate between data and information.

1.2 THE DIFFERENCE BETWEEN DATA AND INFORMATION

Data are widely available; what is scarce is the ability to extract wisdom from them.

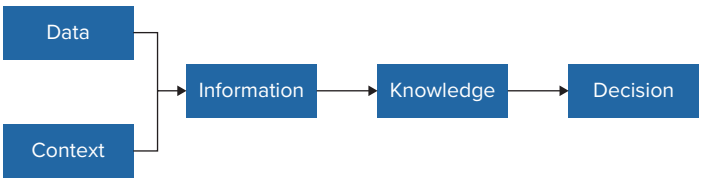
—Hal Varian (UC Berkeley and Chief Economist, Google)

In the previous section, we saw that a vast amount of data is available. But what exactly are data? The term **data** refers to raw numbers and facts that have little meaning on their own. In contrast, **information** is data that are organized in a way that is meaningful to the user in a given context. **Context** is the setting, event, statement, or situation in which the data can be more fully understood and evaluated. In other words, information is data with context.

A brief example will help clarify these key terms. Suppose **Walmart** wants to understand the initial demand for and consumer reaction to a new product, such as the Apple Watch Series 7. Here, context is the seven days immediately after the Apple Watch Series 7 becomes available for sale. The available data are all the individual product reviews on Walmart.com. Information is an organized set of data that helps Walmart assess consumers' reactions to the watch. For example, business analysts may have compiled a list of the words most often used to describe the watch, such as "amazing," "fun," "awesome," "complicated," or "life-changing."

The **information value chain** is composed of the events and processes going all the way from the collection of data to the compilation of information to the ultimate business decision (Exhibit 1.2). To transform data into information, business analysts process the data in some way (for example, by aggregating, sorting, calculating, manipulating, or analyzing it) and then combine it with the appropriate business context (such as time of year, location, or specific business need). Some of the information ultimately becomes knowledge that is helpful in making decisions. **Knowledge** is understanding or familiarity with information gained through learning, and **decisions** are conclusions reached after consideration of the knowledge gained.

Exhibit 1.2
The Information Value Chain



MINI CASE: Tide PODS, Social Media Data, and the Information Value Chain

For an example of the information value chain, consider a data dump of a full year's worth of Instagram posts and Twitter tweets. Without context, these posts and tweets would be data. Now imagine a specific context: You are a business analyst who works for **Procter & Gamble (P&G)**, and the brand manager for Tide (detergent) PODS wants to understand how social media users feel about the product. To provide relevant information to the brand manager, you will need to identify all the Instagram posts and tweets with content related to Tide PODS. Analysis of this information may lead to new knowledge about consumers' preferences regarding product features. The brand manager may then use this knowledge

to make a decision about a proposed new package design. For example, social media users may have objected to the Tide PODS' large plastic container, complaining about its environmental impact. Here, the data may ultimately lead P&G to begin selling Tide PODS in more environmentally conscious packaging.

In addition, P&G may make other decisions based on data gathered from other social media. For example, if social media posts complain about the price of Tide PODS (which are a premium product), managers may decide to develop a budget version of Tide PODS. They may also decide to focus on internet promotions and coupons rather than, say, printed coupons in newspapers, magazines, and mailers.

Decision-Making with Data

Search the Web for social media posts about Tide PODS. Do you see any threads running through the posts, perhaps regarding price, product quality, packaging, or any other aspect of the product? Based on these posts, what recommendations might you make to the brand manager? How would you emphasize the safe use of Tide PODS, given recent headlines about children eating Tide PODS because they look like candy?

PROGRESS CHECK

3. How could a set of tweets about the quality of a new-model **Toyota** RAV4 turn into knowledge capable of affecting decisions at Toyota?
4. How might **Amazon** use the product ratings data on an **Apple** Watch Series 7 to turn data into knowledge? What kind of decisions might this new knowledge help Amazon's managers make?

1.3 SUMMARIZING THE ROLE OF THE BUSINESS ANALYST

LO 1.3

Summarize the role of the business analyst.

Managers and other decision-makers need reliable, relevant information and knowledge to make decisions. As we've seen, business analysts provide this information and knowledge. Often, however, they are not skilled in the advanced statistical and mathematical methods that yield the best, most reliable results. Thus, as Exhibit 1.3 shows, business analysts often serve as intermediaries between decision-makers and data scientists. A **data scientist** is a data specialist who knows how to work with, manipulate, and statistically test data. The business analyst serves as a sort of interpreter, converting the data scientist's technical language into "plain English" for decision-makers.

Business analysts understand the decisions that management must make, but they are also intimately familiar with the characteristics of the data. In addition, they have a working knowledge of data quality, statistical tools, and computer programming—which are all critical to business analysis. Ultimately, the business analyst's job is to bring data-derived knowledge to decision-makers.

Note: Some people prefer the term *liaison* to business analyst, to suggest that the job's key function is to help decision-makers and data scientists communicate with each other.

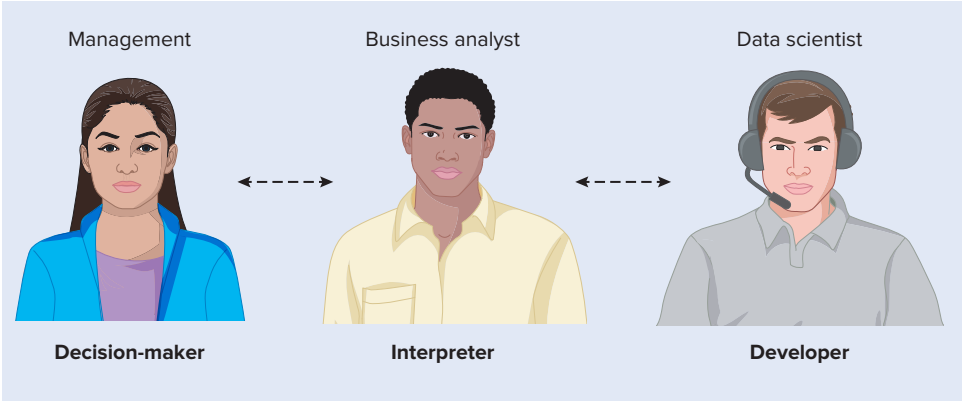


Exhibit 1.3
The Business Analyst Serves as Interpreter Between the Data Scientist and the Decision-Maker

PROGRESS CHECK

5. Who is more likely to know which data will most directly address business questions, a data scientist or a business analyst?
6. If **McGraw Hill**, the publisher of this textbook, wants to determine the most effective ways to market this book, what would the business analyst explain to the data scientist regarding the ideal data sources to consult and potential analysis to perform?

LO 1.4

Describe how the various business functions use business analytics.

1.4 BUSINESS ANALYTICS ACROSS THE DIFFERENT BUSINESS FUNCTIONS

Business analytics refers generally to the use of data to make knowledge, draw conclusions, and address business questions. However, the business analytics for making marketing decisions differs from the business analytics needed to make decisions regarding accounting, finance, and operations. Let's now consider the analytics used in these key business functions.

Marketing focuses on promoting and selling products and services. **Marketing analytics** uses business analytics to measure and improve marketing performance. Perhaps the most important aspect of marketing analytics is providing insights into customer preferences and trends. For example, **Amazon** uses marketing analytics to determine which products to offer, which products to promote/advertise, and which price to charge its customers. Chapter 7 is devoted to defining, explaining, and performing marketing analytics.

Finance is the management of money through investing, borrowing, lending, budgeting, saving, and forecasting. **Financial analytics** uses business analytics to help a company measure, evaluate, and improve its financial performance. Financial analytics also evaluates future investments, such as investments in new equipment, based on risk and expected performance. **American Airlines** uses financial analytics to determine whether it should lease or buy its airplanes, and whether to finance its business with debt or some other means of financing. Chapter 9 is devoted to defining, explaining, and illustrating how financial analytics addresses finance questions.

Operations consists of all the actions needed to run the company and generate income. **Operations analytics** uses business analytics to measure and improve the efficiency and effectiveness of the company's operations. Indeed, achieving one the key goals of any business—"the right product in the right place at the right time"—requires extensive data analysis to ensure product fulfillment through the *supply chain.* Operations analytics also allows the evaluation of a company's human resources (for example, in terms of employee efficiency and turnover), IT operations, sourcing, manufacturing, distribution, and logistics. **Wayfair**, an e-commerce company selling furniture and home goods, uses operations analytics to ensure it has sourced the right furniture products from manufacturers to the Wayfair warehouse, ready for sale and shipping to the final customer. Chapter 10 is devoted to defining, explaining, and providing examples of operations analytics.

Finance, marketing, and operations all generate accounting transactions. *Accounting* measures and records those transactions, and it then communicates the outcomes of those transactions by reporting financial performance. **Accounting analytics** employs business analytics to evaluate financial performance and to address accounting questions related to financial accounting, managerial accounting, auditing, and taxation. For example, during an audit, business analytics are used to detect whether a company committed fraud. **KPMG**, one of the largest accounting firms in the world, uses accounting analytics to evaluate the companies it audits to find errors and look for fraud. In terms of tax planning, accounting analytics helps the company evaluate potential tax deductions to minimize future taxes owed. Chapter 8 is devoted to defining, explaining, and performing accounting analytics.

To emphasize these business functions further, this book provides examples of business analytics in marketing, finance, operations, and accounting in each chapter.

PROGRESS CHECK

7. Which business function would use analytics to suggest the most efficient way to source a car air freshener from Shenzhen, China to a convenience store on Green Street in Champaign, Illinois?
8. Which business function would use analytics to suggest the most efficient way to evaluate whether a company should invest in more efficient manufacturing equipment or a new delivery truck?

1.5 AN INTRODUCTION TO THE SOAR ANALYTICS MODEL

LO 1.5

Identify the components of the SOAR analytics model.

The well-known accounting and advisory firm **EY** argues that all of its existing professionals and new hires need an analytics mindset, which we defined in Section 1.1. In other words, it asks its employees to understand and appreciate the role of business analysts and to develop a similar set of skills. This approach to data is not limited to EY. Businesses across the globe are now asking their employees to develop an analytics mindset.

According to EY, those with an analytics mindset are able to:

- Ask the right questions.
- Extract, transform, and load relevant data.
- Apply appropriate business analytics techniques.
- Interpret and share the results with stakeholders.[4]

[4]Source: Ernst & Young Foundation: E&Y Academic Resource Center (EYARC). 2017. The Analytics Mindset. Available online at http://aaahq.org/Education/Webinars/6-7-17-EY-Academic-Resource-Center-An-Overview-of-Analytics-Mindset-Competencies-and-Case-Offerings.

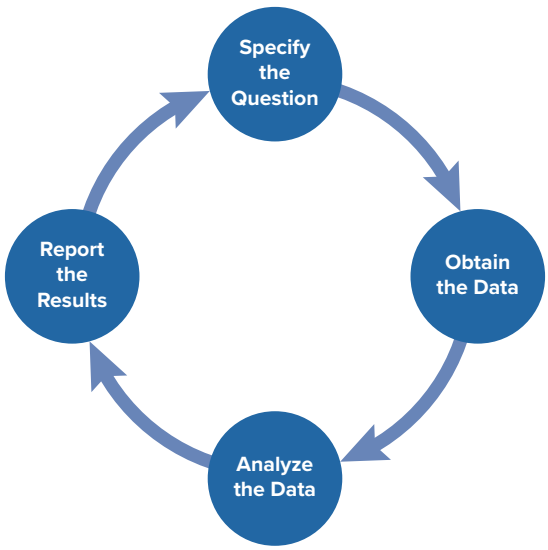
The **SOAR analytics model**, which we use throughout this text, is a useful framework for remembering this set of skills:

1. **Specify the question**: Ask a business question that is answerable with data. (Chapter 1)
2. **Obtain the data**: Gather applicable data and understand the qualities of those data to assess if they can help to answer the business question. (Chapters 2–3)
3. **Analyze the data**: Analyze the data through data manipulation and statistical techniques. (Chapters 4–5)
4. **Report the results**: Communicate the findings to the decision-maker and other interested parties in an effective way. (Chapter 6)

Note that there is a one-to-one correspondence between the characteristics of the analytics mindset and the steps of the SOAR analytics model. Throughout each chapter and each lab activity, we use the SOAR analytics model as a framework to work through the process that begins with specifying the question and ends with reporting the findings to interested parties.

The SOAR model is summarized in Exhibit 1.4, which shows the cyclical nature of the SOAR model. After completing all the steps of the SOAR analytics model, both the analyst and the decision-maker are often more knowledgeable and better able to ask deeper, more refined, questions. Thus, the SOAR analytics model is best viewed as *recursive:* The answer to one question leads to more questions, which lead to more answers and new questions. A good analogy is peeling an onion. After peeling the first layer of the onion, you see the next layer, which you then evaluate and remove to get to the third layer, and so on. Often, as Exhibit 1.4 shows, the business team will iterate through the SOAR analytics model multiple times, narrowing the question (**S**pecify a new question), considering new, potentially better data (**O**btain new data), performing additional analysis (**A**nalyze the new data), and refining the results (**R**eport the latest results) before deciding how to address the issue, problem, or challenge.

Exhibit 1.4
The Recursive SOAR Analytics Model



To understand how the SOAR analytics model works in addressing a specific business question, consider the following example.

Illustration of the SOAR Analytics Model: Predicting Inventory Needs

Imagine that a **Honda** automobile dealership in Tulsa, Oklahoma, is trying to predict the quantity of Honda Pilot SUVs to have available for sale to customers for the following year.

Specify the question.

How many Honda Pilots should we have in inventory for sale to our customers in the Tulsa, Oklahoma, area next year?

Obtain the data.

Gather relevant data:

- How many Honda Pilots were purchased at our dealership last year?
- How many SUVs similar to the Honda Pilot were sold nationwide last year?
- In what year did the last model come out?
- In what year will the next model come out?
- How well is the economy performing overall?

Analyze the data.

- Use data visualizations (including graphs and trend charts) to summarize past sales and evaluate sales trends.
- Use data-driven forecasting to predict Honda Pilot sales at the Tulsa, Oklahoma dealership.
- Predict how sales of the Honda Pilot at the dealership might change under different assumptions, considering best-case, worst-case, and most likely case scenarios.
- Compare sales projections to industry forecasts.

Report the results.

- Determine whether a written report, a visualization (such as a chart or graph), or some combination of text and visuals is the best way to communicate the findings.
- Report the results of the best-case scenario, the worst-case scenario, and the most likely case scenario.
- Evaluate your work: Did the analytics address the question posed? Do decision-makers have what they need to make a decision now? Is additional analysis needed?

In the following sections, we look more closely at each step in the SOAR model.

PROGRESS CHECK

9. What is the difference between obtaining the data and analyzing the data, the second and third steps in the SOAR analytics model?
10. How do the **EY** analytics mindset and the SOAR analytics model sync up with each other?

S: SPECIFY THE QUESTION

The first component of the SOAR analytics model is SPECIFY THE QUESTION. Management has many questions that it needs to answer in order to run the business and make decisions. The more carefully management and analysts specify the question, the more likely we are to gather and use the most appropriate data and perform the right type of analysis.

Here are some examples of questions that decision-makers have asked:

1. **McDonald's**: How is the demographic profile (age, education level, disposable income) changing for the typical customer who eats at our restaurants?
2. **University of Arkansas**: If we purchase new, more efficient copiers, will the long-term savings be sufficient to make up for the expensive price tag?
3. **Michael's**: How long will it take for Christmas products to arrive at our stores if we source the products from Mexico rather than Indonesia?
4. **Walmart**: How will government-imposed tariffs (taxes on imports and exports) affect the sourcing of products to the United States?
5. **Netflix**: How will providing our employees with more comprehensive health insurance affect the level of sales that we need to break even?
6. **Costco**: Which smartphone sold in our Florida stores is the most profitable? Is the most profitable smartphone sold in Florida different than the most profitable smartphone in California?

Generally, the more succinct and specific the question, the better. For example, it is hard to answer a broad question such as "How can we increase our sales?" but easier to address a narrower question such as "How do we sell more Pepsi products in vending machines in Fayetteville, Arkansas?" Narrowing the scope allows us to focus on a specific question and helps us find the necessary data, perform the analysis, and answer the question.

In short, your data won't speak to you unless you ask the right data-analysis questions.[5] We could ask any question in the world, but if we do not have the right data to address the question, we'll never find the answer. Business analysts possess a unique skill set to improve an organization's analytics by asking the right questions.

Note: Although data analytics and business analytics have slightly different meanings, throughout this text we will use the terms *business analysts* and *data analysts* interchangeably.

O: OBTAIN THE DATA

Once data analysts understand the question, they start to consider which data might be most useful in answering it. In their quest for the most useful data, they ask:

- Which types of data need to be collected?
- Which types of data are available?
- What are the sources of useful data?
- How may the data be accessed?
- Will the data adequately address the question?
- Are the data free of errors or inconsistencies?
- Is a particular data set missing a lot of data?
- Will the data have to be adjusted to be usable?
- Are the data biased in some way?
- Have the data been obtained ethically, and will our use of the data protect people's privacy?
- Are the data both relevant and reliable?

Ideally, analysts will use data that are both relevant and reliable. **Relevant data** are directly or closely connected to the question at hand. **Reliable data** reflect the facts or truth with little or no bias. An important component of reliable data is **data integrity**, which refers to the combined accuracy, validity, and consistency of data stored and used over time. Because not all data are both relevant and reliable, analysts need to understand the advantages and disadvantages of using data that are flawed in some way.

[5]M. Durcevic. 2021. "Your Data Won't Speak Unless You Ask It the Right Data Analysis Questions,". *datapine* (January 25). Available online on at https://www.datapine.com/blog/data-analysis-questions/.



When obtaining the data, it is important to consider the ethical implications. According to a 2020 report, "Any consumer with an average number of apps on their phone—anywhere between 40 and 80 apps—will have their data shared with hundreds or perhaps thousands of individuals online." These numbers raise serious questions about data privacy and possible misuses of data.[6] In Chapter 2, we look more closely at the ethical considerations of data collection and data use, and we discuss ethical questions related to data collection and use throughout this book.

Ideally, business analysts want to use data that are well organized and structured, such as a company's financial statements, but they also recognize that disorganized data (for example, trends on Instagram or Twitter) may also be useful.

Sometimes, the company owns the data that it needs, but other times it must buy the data from an external source. In determining whether to acquire and use external data, the company must weigh the cost of use and access against the potential value provided by analyzing the data. Chapter 2 provides more detail on the different data types and sources that can be used to address business questions.

MINI CASE: Hershey Loses Access to Reliable Data to Fulfill Halloween Orders

Due to faulty implementation of a new information system, Hershey could no longer trust its system to provide reliable data. The company was unable to deliver the right amount of candy to its wholesalers, retailers, and customers and, as a result, Hershey lost about $150 million in Halloween candy sales. One manager noted that the company had "a real problem with inventory accuracy, and a lot of the time we didn't have the right inventory to the right place according to our records."[7] Clearly, unreliable data at Hershey had a real impact on the company's operations—and its bottom line.

Decision-Making with Data

How might Hershey have worked around the problem of unreliable data in the short term and the long term?

MINI CASE: Walmart Uses Relevant Data to Identify Trending Products

Walmart leverages social media data to identify trending products to stock in its stores across the world. As just one example, Walmart analyzed social media data to identify the users who seemed to be excited about cake pops, which are bite-sized cakes on a stick. Using this relevant data, Walmart's managers decided to stock cake pops on Walmart's shelves, resulting in incremental sales.[8] Walmart CEO Bill Simon noted, "[Our] ability to pull data together is unmatched."

Decision-Making with Data

Why are data regarding current customer trends considered to be relevant? At a profitable company, how will the revenue due to increased sales compare to the cost of creating, accessing, and using those data?

[6]N. Singer and A. Krolik, "Grindr and OkCupid Spread Personal Details, Study Says," *New York Times*, January 13, 2020, https://www.nytimes.com/2020/01/13/technology/grindr-apps-dating-data-tracking.html. Accessed February 2022.

[7]David F. Carr, "Hershey's Sweet Victory," *Baseline*, December 2002.

[8]Source: https://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109#toc-2. Accessed 8/18/2021.

A: ANALYZE THE DATA

After business analysts have identified the most useful data to address the question, they are ready to perform business analytics. Five common types of analytics are performed, depending on the question to be answered, as summarized in Exhibit 1.5 and discussed below.

Exhibit 1.5
Types of Analytics

Analytics Type	Questions Considered
Descriptive analytics	What happened? What is happening?
Diagnostic analytics	Why did it happen? What are the causes of past results? Why are the results different than expectations?
Predictive analytics	Will it happen in the future? What is the probability something will happen? Can we forecast what will happen?
Prescriptive analytics	What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints?
Adaptive/Autonomous analytics	How can we continuously learn using artificial intelligence (AI)? Can we learn from past and current events with adaptive learning?

1. **Questions such as "What happened?" and "What is happening?" are the basis for *descriptive analytics*.** Such questions may include:
   - What was the average age of our customers last year? Has the average age of our customers changed over the past ten years?
   - How many leads did our marketing campaign generate last quarter?
   - How many of the smartphones that we manufactured were unusable due to flaws in the manufacturing process?
   - How much did we pay in federal taxes last year?
   - What was the shipping time for our products manufactured in Taiwan?
2. **Questions such as "Why did it happen?" and "What are the causes of past results?" and "Why are the results different than expectations?" are the basis of *diagnostic analytics*.** Such questions may include:
   - Why did overall shipping time increase this quarter as compared to last quarter?
   - Why did our overall taxes increase even though our sales did not?
   - Why did labor costs increase last year above expected levels?
   - Did our increased spending on advertising lead to an increase in sales?
3. **Questions such as "Will it happen in the future?" and "What is the probability something will happen?" and "Can we forecast what will happen?" are the basis of *predictive analytics*.** Such questions may include:
   - What will be the demand for our new smartphone next year?
   - Are we likely to be audited by the IRS?
   - Can we forecast stock prices and stock returns?
   - What is the chance that a certain company will go bankrupt?
4. **Questions such as "What should we do, based on what we expect will happen?" and "How do we optimize our performance based on potential constraints?" are the basis of *prescriptive analytics*.** Such questions may include:
   - How can revenues be maximized (or costs be minimized) if world governments increase taxes on imports and exports?
   - How can we optimize our warehouses and transportation to minimize costs?

- Should we rent or lease our headquarters office building?
- Should we make our products in the United States or outsource the manufacturing to another country?

5. **Questions such as "How can we continuously learn using artificial intelligence?" and "Can we learn from past and current events with adaptive learning?" are the basis of *adaptive/autonomous analytics.* Artificial intelligence (AI)** is a collective term for computer systems that can sense their environment, think, learn, and take action in response to what they are sensing and their objectives. AI systems appear to mimic human thought processes.[9] Adaptive/autonomous analytics answers questions such as:
   - How do we continuously update our predictions of needed inventory to more closely meet customer demand?
   - Is our marketing working? How should we adjust our marketing strategies?
   - How do we rebalance and optimize our investment portfolio based on the change in oil prices last week?
   - Based on the decisions of our internal auditing department, could a computer learn which transactions might constitute fraud?

The *analytics ascendency model* in Exhibit 1.6 identifies the skills required for each of the five analytics types. The adaptive/autonomous analytics skill of *rightsight* entails learning the right question to ask your data, at the right time, at the right place, for the right object (or event), in the right context[1] so that actionable information and knowledge can be delivered to the right people in the right form at the right time. We elaborate further on the rightsight skill in Chapter 11.

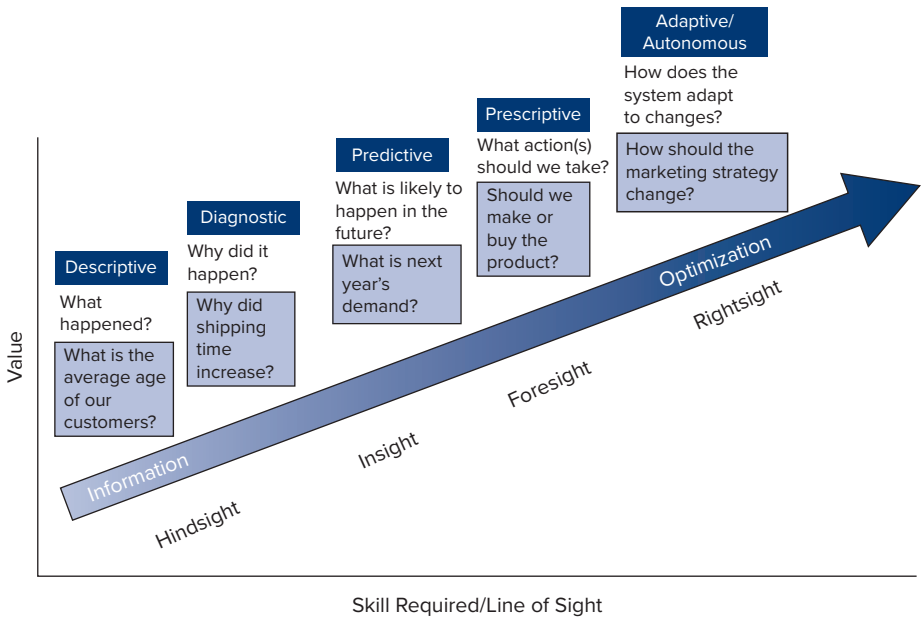


Exhibit 1.6
Analytics Ascendency Model

[9]PwC. 2017. Sizing the Price: What's the Real Value of AI for Your Business and How Can You Capitalise? Available online at https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html. Accessed 8/21/2019.

Once we know the specific question we are trying to answer and obtain the appropriate data, we can perform the correct type of analysis using statistical software tools. Although this text introduces a variety of data analytics software tools, including UiPath, Microsoft Access, Idea/ACL, SAS, SAP Lumira, R, and Python, we focus on the use of Excel, Tableau, and Power BI. These are the three most popular analytics software tools used in the business world, and they are used extensively by business analysts. Learning to work with these tools will prepare you well for your career, whether you are a full-time business analyst, a data scientist, a decision-maker, or just someone who routinely asks and addresses business questions.

R: REPORT THE RESULTS

When the analysis is complete, it is important to communicate and report the results to the decision-makers. This report may take the form of a written report, statistical results, bar charts and other graphics, or any combination of these. The results might be reported just once in a **static report** that is not constantly updated. Or the report may undergo continuous, real-time updates (a **dynamic report**) delivered through a **dashboard**, which is a graphical summary of various measures continuously tracked by a company.



SOARing to Success

Using Software for Success

We use Excel, Tableau, and Power BI as our analytics tools in this text. I find that most internships and entry-level business analyst positions require some proficiency in Excel, along with some skill in a visualization tool such as Tableau or Power BI. Some businesses prefer one software tool over another, but in general the problems and labs throughout this text will teach you an approach to accessing and analyzing data. While you may not use the tools you know best at your first or next job, the approach to accessing and analyzing data will serve you throughout your career.

Need an introduction or a refresher on these tools? Appendices A, B, and C provide tutorials on Excel, Tableau, and Power BI, respectively.

MINI CASE: Netflix Wants to Improve Its CineMatch Movie Recommendation System[10]

From 2006 to 2009, Netflix ran a contest asking contestants to design software that would improve CineMatch, its famous in-house recommender system, by at least 10 percent. It promised a $1 million prize to the winner. We can apply the SOAR analytics model to this contest and its outcome.

- **Specify the question**. Netflix asked, "How can we improve our movie recommendation system for users?"
- **Obtain the data**. Netflix provided contestants with a data set composed of 100,480,507 ratings that 480,189 Netflix subscribers gave to 17,770 movies.
- **Analyze the data**. The contestants used various data analysis methods to make recommendations for improving the CineMatch system. The winner was BellKor's Pragmatic Chaos algorithm. (An *algorithm* is a process or set of rules to be followed in computer calculations.)

[10]Source: https://towardsdatascience.com/netflix-recommender-system-a-big-data-case-study-19cfa6d56ff5. Accessed 8/18/2021.

- **Report the results**. The new algorithm now regularly makes more than 5 billion movie recommendations to its customers each year. Netflix received reports on whether this recommendation system actually improved Netflix customer satisfaction and retention.[11]
- **Recursive SOAR analytics model**: The algorithm continues to be tweaked to further improve customer satisfaction and retention.

Decision-Making with Data

To continue improving the satisfaction of the Netflix subscriber base, what questions might you ask? What types of data might you gather to help you answer these questions?

PROGRESS CHECK

11. "What are sales expected to be next year?" is which type of question?
    a. What happened? What is happening?
    b. Why did it happen? What are the causes of past results?
    c. Will it happen in the future? What is the probability something will happen? Can we forecast what will happen?
    d. What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints?
    e. Can we learn from past and current events with adaptive learning?
12. Refer to the answer choices in question 11 above. "What is the monthly subscription price that optimizes income for **Netflix**?" is which type of question?
13. Why is the SOAR analytics model recursive? Is the analysis ever complete?

1.6 USING EXPLORATORY AND EXPLANATORY DATA VISUALIZATIONS IN BUSINESS ANALYTICS

LO 1.6

Describe the use of exploratory data visualizations and explanatory data visualizations.



DATA VISUALIZATION

One key analytics tool used in the ANALYZE THE DATA and REPORT THE RESULTS steps of the SOAR analytics model is the use of visualizations. A **data visualization** is a graphic representation of data, often in the form of a graph, chart, or other image. *Exploratory* data visualizations can reveal patterns in the data, which can lead to greater insight. *Explanatory* data visualizations are very useful for communicating the findings of the business analysis to stakeholders.[12]

Exploratory Visualizations

An **exploratory visualization** is a graphical representation that is useful for uncovering patterns and useful insights in the data, generally as part of descriptive or diagnostic analytics. These visualizations are typically used during the ANALYZE THE DATA step of the

[11] Shiv Mehta, "Netflix: How Businesses Can Learn Management Innovation," October 13, 2018, https://www.linkedin.com/pulse/netflix-how-businesses-can-learn-management-shiv-mehta-fmva-/?articleId=6452774975547015168.

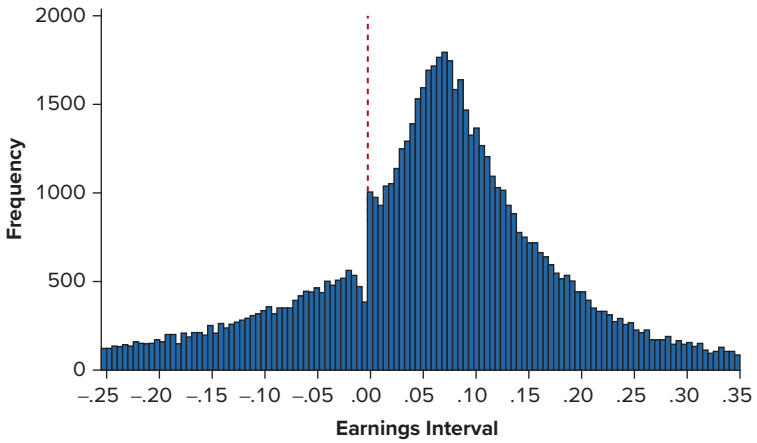
[12] Ann Dzuranin, "Explanatory Data Visualizations," *Strategic Finance,* January 1, 2022, https://sfmagazine.com/post-entry/january-2022-explanatory-data-visualizations/. Accessed January 7, 2022.

SOAR model to explore and understand the data. For example, when performing descriptive or diagnostic analytics to determine what happened, a business analyst may decide to graph the data to gain a better understanding of what is occurring.

For example, suppose that an analyst wants to understand how profitability is distributed over time across all the firms trading on a U.S. public exchange. The analyst might choose to plot the return on equity (ROE, which equals net income divided by stockholders' equity) at each earnings interval across all of those firms. Accounting researchers David Burgstahler and Ilia Dichev[13] did exactly that, creating the visualization shown in Exhibit 1.7 for the period from 1976 to 1994. This exploratory visualization shows what earnings were reported over the time period in each earnings (or profitability) interval, so it is a type of descriptive analytics. The bar at each value on the earnings interval shows how many firms fell in that interval (frequency).

Exhibit 1.7
Distribution of Return on the Market Value of Stockholders' Equity, 1976–1994

Source: David Burgstahler and Ilia Dichev, "Earnings Management to Avoid Earnings Decreases and Losses," *Journal of Accounting and Economics* 24, no. 1 (1997).



For the most part, the visualization displays the expected bell-shaped distribution of earnings that often shows up in stock market returns. However, the visualization shows an anomaly—in this case, a discontinuity, right around zero. Just below zero, there seem to be missing observations or a lower-than-expected frequency, suggesting some type of an anomaly. And just above zero, there seem to be extra observations (or a higher-than-expected frequency), again suggesting an anomaly.

This visualization will likely lead to additional questions:

- What explains the differences in profitability level from company to company?
- Are managers responsible for the higher-earning companies' high profit levels?
- Do managers use the flexibility within accounting standards to make earnings higher (or lower) than the expected? If they do manage earnings, how do they do that?
- What incentives do managers have to manage earnings up or down?
- Do managers manage earnings to maximize their own bonuses?

Explanatory Visualizations

As previously noted, the data analyst serves as an interpreter, or liaison, between the data scientist and decision-makers. Indeed, a business analyst should be particularly good at reporting the results to decision-makers and all other interested parties. Very often the

[13]David Burgstahler and Ilia Dichev, "Earnings Management to Avoid Earnings Decreases and Losses," *Journal of Accounting and Economics* 24, no. 1 (1997), pp. 99–126.

report includes **explanatory visualizations** that are useful in communicating the findings of the analysis to stakeholders.

Increasingly, data visualizations are preferred to written content to communicate results. Why? Studies have shown that 91 percent of people prefer visual content over written content.[14] For example, on Facebook, photos have an interaction rate of 87 percent, compared to 4 percent or less for other types of posts, such as links or text.[15] In addition, the brain processes images 60,000 times faster than it processes text, and 90 percent of information transmitted to the brain is visual.[16]

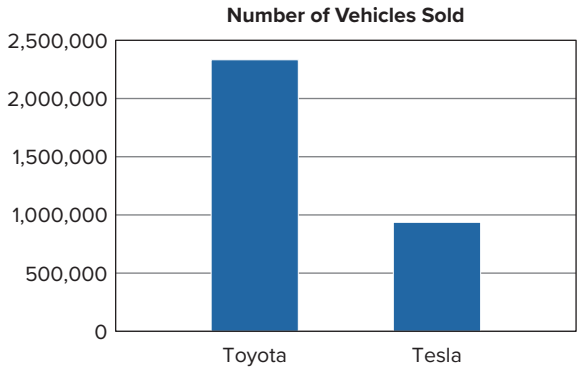
Different visualizations are used for different purposes. Exhibit 1.8 summarizes some common purposes for data visualization as well as five corresponding business questions and common visualization types: bar charts, pie charts, histograms, line graphs, and scatterplots. Exhibit 1.9 shows samples of these five basic visualization types. Chapter 6 further describes the purpose of each data visualization type and introduces other useful types of visualizations. In addition, you will cover of data visualization, along with data visualization exercises, in each chapter and in many of the labs.

Exhibit 1.8
Purposes of Data Visualization, Business Examples, and Visualization Types

Purpose of Visualization	Business Examples	Common Visualization Type (See Exhibit 1.9)
Comparison of values	Which company sold more cars last year, **Toyota** or **Tesla**?	Column chart (Panel A) or Bar chart
Composition of values	What proportion of sales comes from each of the five divisions of the **J.B. Hunt** company?	Pie chart (Panel B)
Distribution of values	How many days on average did it take **Wayfair** to fulfill a set of 20,000 orders?	Histogram (Panel C)
Trends of values over time	How has **Google**'s ad revenue changed over the last five years?	Line graph (Panel D)
Relationships between values	What is the relationship between advertising expense and sales revenue at **Costco**?	Scatterplot (Panel E)

Exhibit 1.9
Common Types of Data Visualization

(A) Column chart to compare values



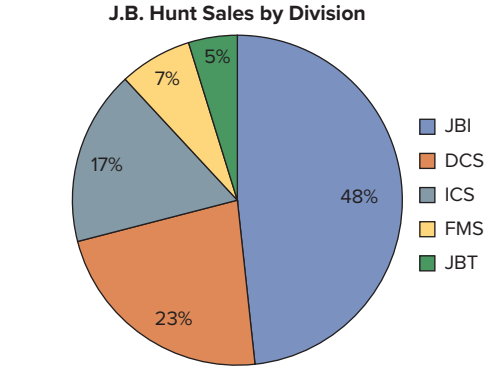
[14]Zohar Dayan, "Visual Content: The Future of Storytelling," *Forbes*, April 2, 2018, https://www.forbes.com/sites/forbestechcouncil/2018/04/02/visual-content-the-future-of-storytelling/?sh=6517bfbe3a46. Accessed 1/14/2021.

[15]Hannah Whiteoak, "Six Reasons to Embrace Visual Commerce in 2018," *Pixlee*, https://www.pixlee.com/blog/six-reasons-to-embrace-visual-commerce-in-2018/. Accessed 3/4/2022.

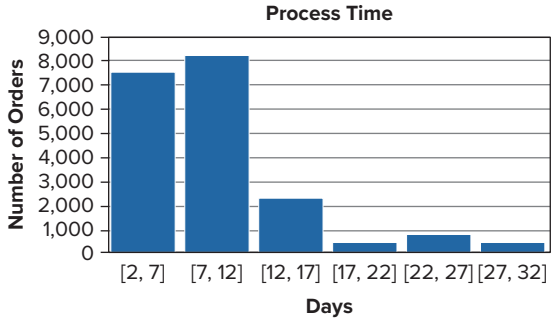
[16]Harry Eisenberg, "Humans Process Visual Data Better," *Thermopylae Sciences + Technology,* 9/15/2014, https://www.t-sciences.com/news/humans-process-visual-data-better#:~:text=Visualization% 20works%20from%20a%20human,to%20the%20brain%20is%20visual. Accessed 1/14/2021.

Exhibit 1.9 (*Continued*)

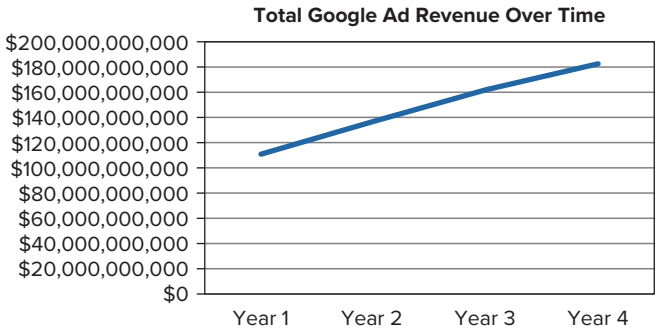
(B) Pie chart to visualize composition of values



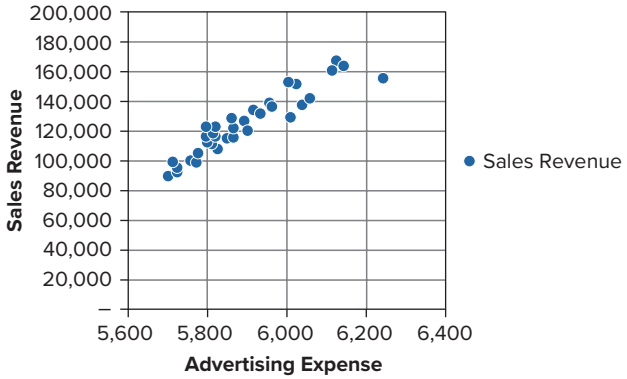
(C) Histogram to show the distribution of values (number of days to process orders)



(D) Line graph to show trend (total Google ad revenue over time)



(E) Scatterplot to show relationship between values (advertising expense and sales revenue)



LAB CONNECTIONS

There are three labs at the end of this chapter. Each exemplifies the use of the entire SOAR analytics model from specifying a question to reporting the results. Each lab asks you to use at least one software tool: Excel, Tableau, or Power BI. Completing the labs is an essential component of this course. They will provide you with the software skills that employers are looking for. They will help you answer questions such as:

- Does a potential business opportunity pay enough money for you to leave your current job? (Lab 1.1 Excel)
- Who are your most profitable customers? (Lab 1.2 Excel, Lab 1.2 Tableau, and Lab 1.2 Power BI)
- What are your most profitable products? (Lab 1.3 Excel, Lab 1.3 Tableau, and Lab 1.3 Power BI)

PROGRESS CHECK

14. In which step of the SOAR model are exploratory visualizations typically created? Which step uses explanatory visualizations?
15. If a company would like to track actual daily sales as compared to its sales targets, should it use a dynamic report or a static report? Explain.

Chapter Summary

Businesses exist to create value. Managers call on business analysts to answer questions that may help them create value. In this chapter, we discussed the huge, and increasing, volume of data available, which gives rise to the role of the business analyst. We learned that data with context can create valuable information that can turn into knowledge and can impact business decisions. We saw that business analytics can address different types of questions in the different business functions (marketing, accounting, finance, and operations). We also summarized the SOAR analytics model as a framework for performing business analytics, discussed the type and scope of questions that business analysts ask as the first component of the SOAR model, and introduced the importance of data visualization.

Key Terms

accounting analytics (*9*) Business analytics to evaluate financial performance and to address accounting questions.

analytics mindset (*5, 9*) The willingness and ability to specify which questions need to be addressed, find and extract pertinent data that might address those questions, analyze those data, and then report the results of business analytics to decision-makers.

artificial intelligence (AI) (*15*) A collective term for computer systems that can sense their environment, think, learn, and take action in response to what they are sensing and their objectives.

business analyst (*4*) A data specialist who curates and uses data to help an organization make effective business decisions.

business analytics (*8*) The use of data to make knowledge, draw conclusions, and address business questions.

business process (*4*) A coordinated, standardized set of activities conducted by both people and equipment to accomplish a specific business task.

business value (*4*) All the items, events, and interactions that determine a company's financial health.

context (*6*) The setting, event, statement, or situation in which data can be more fully understood and evaluated.

dashboard (*16*) A graphical summary of various measures tracked by a company.

data (*6*) Raw numbers and facts that have little meaning on their own.

data integrity (*12*) The combined accuracy, validity, and consistency of data stored and used over time.

data overload (*5*) Access or exposure to too much data; prevents data from being properly synthesized and interpreted.

data scientist (*7*) A data specialist who knows how to work with, manipulate, and statistically test data.

data visualization (*17*) The graphic representation of data, often in the form of a graph, chart, or other image.

decision (*6*) A conclusion reached after considering the knowledge gained.

dynamic report (*16*) A report characterized by constant update, change, or activity.

explanatory visualization (*19*) A graphical representation useful in communicating the findings of the analysis to stakeholders.

exploratory visualization (*17*) A graphical representation that is useful for uncovering patterns and useful insights in the data, generally as part of descriptive or diagnostic analytics.

financial analytics (*8*) Business analytics to measure, evaluate, and improve financial performance.

information (*6*) Data organized in a way that is meaningful to the user in a given context.

information value chain (*6*) The events and processes from the collection of data to the compilation of information to an ultimate business decision.

knowledge (*6*) Understanding of or familiarity with information gained through learning.

marketing analytics (*8*) Business analytics to measure and improve a company's marketing performance.

operations analytics (*9*) Business analytics to measure and improve the efficiency and effectiveness of a company's operations.

relevant data (*12*) Data that are directly or closely connected to the question at hand.

reliable data (*12*) Data that reflect the facts or truth with little or no bias.

SOAR analytics model (*10*) A framework for performing business analytics, which includes four steps: **S**pecify the Question, **O**btain the Data, **A**nalyze the Data, and **R**eport the Results.

static report (*16*) A one-time report that remains unchanged.

ANSWERS TO PROGRESS CHECKS

1. The increasing availability of data has given rise to the role of the business analyst. More data often means the potential for more data-driven knowledge and insights. However, the increasing amount of data may also lead to data overload, which may overwhelm the analyst. Recognizing the value of data without becoming overwhelmed by it requires careful balance.
2. Business analysts are trained to have an analytics mindset. They understand how to access, synthesize, analyze, and interpret data. Because data analysis is the main focus of their job, they do not have a lot of other tasks competing for their time. They can focus intensely on working with data to help managers answer their questions and identify possible solutions.
3. Tweets might express both positive and negative sentiments and attitudes about the latest **Toyota** RAV4 model. These tweets might affect the way the car is advertised, marketed, and priced. For example, the tweets might reveal that the model is popular with families with young children but not popular with recent college graduates. Here, analyzing tweets is part of the information value chain, which combines data with context to produce knowledge that can help managers with decision-making.
4. **Amazon**'s product ratings data on the **Apple** watch could be combined with sales data to help Amazon determine past demand and forecast future demand, how many Apple Watches to hold in stock, and what consumers like and don't like about the watch. Of course, the data could be used in many other ways, too.
5. A business analyst is more likely to know which data are needed to address business questions. The business analyst is likely to understand both the decision-maker's information needs and the data scientist's statistical techniques.
6. The business analyst will explain that there are many different means of marketing a textbook (for example, to professors, to students directly, Web-based ads, direct sales, displays at academic conventions), and each method may have a different level of effectiveness. The **McGraw Hill** business analyst can work with the data scientist to discover which data sources are readily available and will provide the best information and knowledge for decision-making. They could then work together to analyze the data and report the results.
7. Operations analytics. In this example, the manufacturer of the air freshener is concerned with an efficient *supply chain,* which involves all aspects of manufacturing, distribution, and logistics. The supply chain is an essential component of the company's operations.
8. Financial analytics uses business analytics to help a company measure and evaluate its financial performance and its investments in new equipment, such as more efficient manufacturing equipment or a new delivery truck.
9. Obtain the data, the second step in the SOAR analytics model, involves gathering applicable data to answer the question. Analyzing the data, the third step in the SOAR analytics model, involves business analysis through data manipulation and statistical techniques.
10. The linkages between the **EY** analytics mindset and the SOAR analytics model are as follows:

EY Analytics Mindset	**SOAR Analytics Model**
Ask the right questions.	Specify the question.
Extract, transform, and load relevant data.	Obtain the data.
Apply appropriate business analytics techniques.	Analyze the data.
Interpret and share the results with stakeholders.	Report the results.

11. (c). Forecasting sales next year requires an answer to the question "Will it happen in the future?" Finding the answer requires predictive analytics. In contrast, "What were sales last year?" is an example of the question "What happened?" and requires descriptive analytics.
12. Optimizing income or other outcomes at **Netflix** requires an answer to the questions "What should we do, based on what we expect will happen?" and "How do we optimize our performance based on potential constraints?" and the use of prescriptive analytics.
13. The SOAR analytics model is recursive or cyclical because the answer to one question leads to more questions, which lead to more answers and new questions. The recurring nature of the SOAR model allows the business team to narrow its questions and improve the quality of the answers.
14. While exploratory visualizations help data analysts discover insights and patterns in the data (as part of the ANALYZE THE DATA step in the SOAR model), explanatory visualizations communicate the insights and findings of the analytics to stakeholders.
15. Tracking daily sales would require almost continuous updates, which require dynamic reports, rather than static reports, which do not change.

connect

Multiple-Choice Questions

1. (LO 1.1) A coordinated, standardized set of activities conducted by both people and equipment to accomplish a specific business task is called a:
   a. business process.
   b. business analysis.
   c. business procedure.
   d. business value.
2. (LO 1.2) According to the information value chain, data combined with context is:
   a. information.
   b. knowledge.
   c. insight.
   d. value.
3. (LO 1.5) Which phase of the SOAR analytics model addresses the proper way to communicate results to the decision-maker?
   a. Specify the question
   b. Obtain the data
   c. Analyze the data
   d. Report the results
4. (LO 1.5) Which phase of the SOAR analytics model involves finding the most appropriate data needed to address the business question?
   a. Specify the question
   b. Obtain the data
   c. Analyze the data
   d. Report the results

5. (LO 1.5) Which questions seek information about Tesla's sales in the next quarter?
   a. What happened? What is happening?
   b. Why did it happen? What are the causes of past results?
   c. Will it happen in the future? What is the probability something will happen? Can we forecast what will happen?
   d. What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints?
6. (LO 1.5) Which questions seek information on the routing of products from Queretaro, Mexico to Chicago, United States in the last quarter?
   a. What happened? What is happening?
   b. Why did it happen? What are the causes of past results?
   c. Will it happen in the future? What is the probability something will happen? Can we forecast what will happen?
   d. What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints?
7. (LO 1.5) Which questions ask why net income is increasing when revenues are decreasing, counter to expectations?
   a. What happened? What is happening?
   b. Why did it happen? What are the causes of past results?
   c. Will it happen in the future? What is the probability something will happen? Can we forecast what will happen?
   d. What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints?
8. (LO 1.5) Which questions help managers understand how to organize future shipments based on expected demand?
   a. What happened? What is happening?
   b. Why did it happen? What are the causes of past results?
   c. Will it happen in the future? What is the probability something will happen? Can we forecast what will happen?
   d. What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints?
9. (LO 1.5) Which term refers to the combined accuracy, validity, and consistency of data stored and used over time?
   a. Data integrity
   b. Data overload
   c. Data value
   d. Information value
10. (LO 1.3) A specialist who knows how to work with, manipulate, and statistically test data is a:
    a. decision maker.
    b. data scientist.
    c. data analyst.
    d. decision scientist.
11. (LO 1.4) Which type of analyst predicts the amount of money that a company will receive from its customers to help management evaluate future investments based on expected investment performance, such as investments in equipment or employee training?
    a. Marketing analyst
    b. Operations analyst
    c. Financial analyst
    d. Accounting analyst

12. (LO 1.4) Which type of analyst addresses questions regarding taxes and auditing?
    a. Marketing analyst
    b. Operations analyst
    c. Financial analyst
    d. Accounting analyst
13. (LO 1.5) Suppose a company has timely product reviews that are available when needed, but the reviews are biased. These product reviews are which type of data?
    a. Reliable
    b. Relevant
    c. Curated
    d. Consistent
14. (LO 1.6) Which common visualization type shows trends in values over time?
    a. Line graph
    b. Scatterplot
    c. Pie chart
    d. Bar chart
15. (LO 1.6) Which common visualization type shows the composition of values over time?
    a. Line graph
    b. Scatterplot
    c. Pie chart
    d. Bar chart

Discussion Questions

1. (LO 1.1) Give five examples of business processes at **Tesla**. How do they create business value for Tesla and its shareholders?
2. (LO 1.2) Explain the information value chain by summarizing how data are transformed into knowledge insights for decision-making. Use the example of a book review on **Amazon** and how it might lead Amazon to decide how many of those books to stock at its warehouses.
3. (LO 1.3) Explain the information value chain by summarizing how data are transformed into knowledge insights for decision-making. Use the example of a book review of this book on **Amazon** and how it might help the publisher, **McGraw Hill**, determine whether to revise this book for a new, updated edition as the discipline of data analytics evolves.
4. (LO 1.3) Explain the difference between a decision-maker, a data scientist, and a business analyst. What is the role of each?
5. (LO 1.4) Compare and contrast marketing analytics with accounting analytics. How are they similar? How are they different?
6. (LO 1.4) Compare and contrast financial analytics with operations analytics. How are they similar? How are they different?
7. (LO 1.5) Identify the four steps in the SOAR analytics model. Explain how marketing analysts might use the SOAR model to help **Netflix** better understand its customers.
8. (LO 1.5) Identify the five sets of questions that are asked in business analytics. Brainstorm different types of business-related questions that might fit in each of the five question sets.
9. (LO 1.5) Consider the data qualities of relevance and reliability. Explain why each quality and both qualities together are valuable to business analysts.
10. (LO 1.6) Some argue that a picture is worth a thousand clicks. Argue for and against the use of data visualizations when reporting the results.

connect

Brief Exercises

1. (LO 1.1, 1.2, 1.4, 1.5) Match each business analytics term to its definition.

Term	Definition
1. Information	a. The setting, event, statement, or situation in which data can be more fully understood and evaluated
2. Dashboard	b. Access or exposure to too much data, which prevents data from being properly synthesized and interpreted
3. Dynamic report	c. A graphical summary of various measures tracked by a company
4. Data overload	d. A report characterized by constant update, change, or activity
5. Context	e. Data organized in a way that is meaningful to the user

2. (LO 1.5) For each business analytics task, identify the correct step in the SOAR analytics model.

Business Analytics Task	SOAR Analytics Model Component (Specify the question, Obtain the data, Analyze the data, or Report the results)
Creating a dashboard that provides daily sales in the Northeast region of the country	
Trading off accessing cleaner data (fewer errors) with a data set that is less expensive	
Small business owner trying to decide which products to sell on the **eBay** platform	
Using a statistical test to see if customer characteristics changed from one year to the next	
Deciding whether to use static or dynamic reports to communicate findings to management	

3. (LO 1.5) For each business analytics task, identify the correct step in the SOAR analytics model.

Business Analytics Task	SOAR Analytics Model Component (Specify the question, Obtain the data, Analyze the data, or Report the results)
Deciding which data to use in the analysis	
Running a statistical analysis to evaluate the impact of advertising on sales of targeted products	
Extracting data from blogs to analyze consumer sentiment regarding **Verizon** wireless service	
Communicating current customers' key characteristics to management	
Analyzing how profits will change if taxes on Chinese-sourced products increase in the coming year	
Asking which vendor is most reliable based on past performance	

4. (LO 1.1, 1.2, 1.4, 1.5) Match each business analytics term to its definition.

Term	Definition
1. Data integrity	a. Raw numbers and facts that have little meaning on their own
2. Business analytics	b. Analytics used to evaluate the promotion, distribution, and selling of a product or service.
3. Data	c. The combined accuracy, validity, and consistency of data stored and used over time
4. Accounting analytics	d. The use of data to make knowledge, draw conclusions, and address business questions
5. Marketing analytics	e. Analytics used to evaluate the financial performance of a business.

5. (LO 1.5) For each business analytics question below, identify the questions it is answering. Choose from this list:
   a. What happened? What is happening?
   b. Why did it happen? What are the causes of past results?
   c. Will it happen in the future? What is the probability something will happen? Can we forecast what will happen?
   d. What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints?
   e. How can we continuously learn using artificial intelligence? Can we learn from past and current events with adaptive learning?

Business Analytics Question	Question Type (Indicate letter a, b, c, d, or e)
What are the key characteristics of the customers who purchased from us last year?	
Why did **Delta Airlines**' on-time departures improve this past month?	
What is the expected return on an investment in stocks instead of bonds, based on what we expect will happen?	
Can we predict how long it will take our vendor to get its products to our distribution center?	
Why did our company's import taxes increase last year?	
What are the expected sales of bananas next year at the **Kroger** chain of grocery stores?	
Could a computer learn from decisions made by inventory managers in order to auto-order supplies whose stock is dwindling?	

6. (LO 1.2) Match each definition to the correct term. All terms refer to the information value chain.
   a. Data
   b. Context
   c. Information
   d. Knowledge
   e. Decision

Definition	Term
Raw numbers and facts that have little meaning on their own	
Understanding of, or familiarity with, information gained through learning	
The setting, event, statement, or situation in which data can be more fully understood and evaluated	
A conclusion reached after considering the knowledge gained	
Data organized in a way that is meaningful to the user in a given context	

Problems

1. (LO 1.4, 1.5) How might data analytics be used in the each of the following situations? Explain your answer.
   a. Determining where to locate a **Black Bear Diner** (a restaurant that serves homestyle and old-fashioned comfort foods) in Lawrence, Kansas
   b. Deciding which gaming app to build and market for Android phones
   c. Determining the appropriate venue for a Taylor Swift concert in College Station, Texas
   d. Analyzing the depth of a nursing shortage at the regional hospital in Fayetteville, Arkansas, along with possible solutions
2. (LO 1.4, 1.5) **Tesla** is considering each of the following questions. Identify whether the company will use descriptive, diagnostic, predictive, prescriptive, or adaptive/autonomous tools to answer them. Provide support for your answers.
   a. What are the forecasted Tesla sales in China for the next three to five years?
   b. If the costs of manufacturing increase due to increased regulation in the United Kingdom, how will profit be affected?
   c. What taxes did Tesla pay in Malaysia in the past four quarters?
   d. What was the return on investment for Tesla stockholders over the past five years?
   e. How should we adjust our R&D (research and development) based on expected improvements in battery technology?
   f. Why were materials more expensive than anticipated?
3. (LO 1.4) For each of the questions in Problem 2, determine which type of analytics might address it: marketing analytics, accounting analytics, financial analytics, and/or operations analytics. Some questions may have two possible answers. Support your reasons for each answer.
4. (LO 1.2) Explain the **Netflix** mini case in this chapter and its enhanced recommender system in terms of the information value chain. Specifically, how did Netflix turn data and context into decisions?
5. (LO 1.5) The **Harps** grocery store chain buys its paper goods (paper towels, toilet paper, tissues) in bulk from one supplier to get a discount. Assume that about once a month it runs out of stock of a few of these items. Harps is concerned that these stockouts are causing customers to shop elsewhere, and it is attempting to understand whether the issues are in its purchasing procedures or its supply chain. Use the steps in the SOAR analytics model to specify Harps' question and to suggest how its business analysts will obtain the data, analyze the data, and report the results.
6. (LO 1.5) The recommended level of retirement savings an individual should have saved at each age as follows:
   a. At age 30: 1X income at age 30
   b. At age 40: 3X income at age 40
   c. At age 50: 5X income at age 50
   d. At age 70: 7X income at age 70

Bob makes $90,000 a year in income at age 50, and his salary is growing at 5% per year. He currently has retirement savings of $400,000. He wants to know if he will have enough income to retire at age 70. What questions could he ask regarding his retirement savings that might be addressed using descriptive, diagnostic, predictive, and prescriptive analytics? What assumptions are being made along the way to help with his analysis?

7. (LO 1.5) Download the DJI Mavic Drone SKU data set. In Excel, calculate the gross margin (Sales Price – Cost) and gross margin percentage [(Sales Price – Cost)/Sales Price] for each SKU (product). (This problem may require a review of Excel skills, which you will find in Appendix A at the back of this book.)
   a. Which product SKU and description has the highest gross margin? (*Hint:* Use =MAX() Excel function to calculate.)
   b. Which product SKU and description has the highest gross margin percentage? (*Hint:* Use =MAX() Excel function to calculate.)
   c. Which product SKU and description has the lowest gross margin? (*Hint:* Use =MIN() Excel function to calculate.)
   d. Which product SKU and description has the lowest gross margin percentage? (*Hint:* Use =MIN() Excel function to calculate.)
   e. What are the average and median gross margin? (*Hint:* Use =AVERAGE() and =MEDIAN() Excel functions to calculate.)
8. (LO 1.5) Download the SkyDio Drone SKU data set. In Excel, calculate the gross margin (Sales Price – Cost) and gross margin percentage [(Sales Price – Cost)/Sales Price] for each SKU (product). (This problem may require a review of Excel skills, which you will find in Appendix A at the back of this book.)
   a. Which product and SKU description has the highest gross margin? (*Hint:* Use =MAX () Excel function to calculate.)
   b. Which product and SKU description has the highest gross margin percentage? (*Hint:* Use =MAX () Excel function to calculate.)
   c. Which product and SKU description has the lowest gross margin? (*Hint:* Use =MIN() Excel function to calculate.)
   d. What is the gross margin percentage of the product SKU and description with the lowest gross margin percentage? (*Hint:* Use =MIN() Excel function to calculate.)
   e. What are the average and median gross margin percentage (round to three digits)? (*Hint:* Use =AVERAGE() and =MEDIAN() Excel functions to calculate.)
9. (LO 1.5) Download the file Monthly Sales Dataset.xlsx. Using Excel, summarize monthly sales and compute the percent change from one month to the next. To calculate percent change, calculate (Current Month Sales – Prior Month Sales)/Prior Month Sales. (This problem may require a review of Excel skills, which you will find in Appendix A at the back of this book.)

Required

   a. Use a pivot table to sum monthly sales.
      i. What is the total amount of August monthly sales?
      ii. What is the total amount of December monthly sales?
      iii. Capture a screenshot of the pivot table and label it Submission 1 in a Word file titled Monthly Sales Submission.docx.
   b. Modify the value field settings for the pivot table (as shown in Appendix A) and calculate the percent change from one month as compared to the previous month.
      i. What is the percentage change in monthly sales from June to July? (Round to 2 decimal points.)
      ii. What is the percentage change in monthly sales from November to December? (Round to 2 decimal points.)
      iii. Capture a screenshot and label it Submission 2 in a Word file called Problem 1.9

10. (LO 1.5) Download the file Regional Sales Dataset.xlsx. Using Excel, summarize sales by month, by sales region (regions 1–4), and by both month and region using a pivot table. (This problem may require a review of Excel skills, which you will find in Appendix A at the back of this book.)

Required

a. Use a pivot table to sum sales by month.
   i. What are the January monthly sales?
   ii. What are the February monthly sales?
   iii. Capture a screenshot of the pivot table and label it Submission 1 in a Word file titled Regional Sales Submission.docx.

b. Use a pivot table to sum sales by region.
   i. What are the total sales for region 2?
   ii. What are the total sales for region 4?
   iii. Capture a screenshot of the pivot table and label it Submission 2 in a Word file titled Regional Sales Submission.docx.

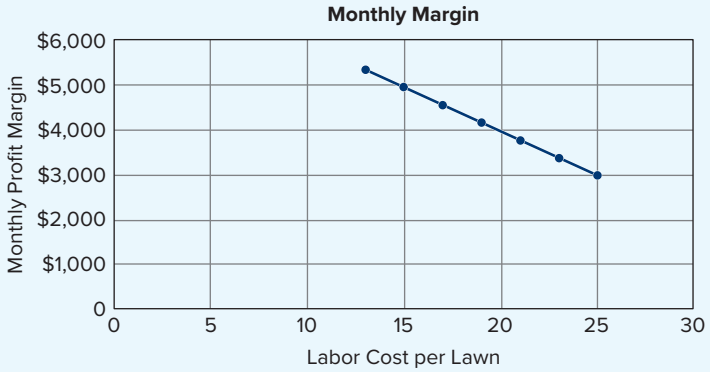
c. Use a pivot table to sum sales by both month and region. (*Hint:* In the pivot table, put month in the rows and region in the columns.)
   i. What are the December sales for region 1?
   ii. What are the April sales for region 4?
   iii. Capture a screenshot of the pivot table and label it Submission 3 in a Word file titled Regional Sales Submission.docx. Upload Regional Sales Submission.docx, if it is assigned.

Data Visualization Exercises



DATA VISUALIZATION

1. (LO 1.6) What does the following data visualization tell us about the relationship between labor cost per lawn mowed and monthly (profit) margin (the extent to which revenues exceed expenses) as shown on the vertical axis?



2. (LO 1.6) Is the data visualization shown in Exercise 1 better or worse than the same data in the following tabular format? Explain.

Labor Cost per Lawn	Monthly Margin
13	$5,350
15	$4,950
17	$4,550
19	$4,150
21	$3,750
23	$3,350
25	$2,950

CHAPTER 1 LABS

LAB 1.1 **Excel:** Assessing Profitability of a Mowing Business
LAB 1.2 **Excel:** Assessing Customer Profitability
LAB 1.2 **Tableau:** Assessing Customer Profitability
LAB 1.2 **Power BI:** Assessing Customer Profitability
LAB 1.3 **Excel:** Assessing Product (SKU) Profitability
LAB 1.3 **Tableau:** Assessing Product (SKU) Profitability
LAB 1.3 **Power BI:** Assessing Product (SKU) Profitability

Lab 1.1 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Assessing Profitability of a Mowing Business

Keywords

Practicing Excel Skills, Profitability

Decision-Making Context

A friend is retiring from a lawn-mowing business and is asking if you would like to take over his business. He is not trying to make money on the sale of the business. Instead, he is concerned with (1) taking care of his existing customers and (2) having someone take over existing contracts on equipment he has purchased.

He currently mows 50 lawns each week and charges $50/lawn. He mows lawns himself and has two people working for him. He supplies all equipment:

> **Truck:** He has one truck that he purchased. The amount he still owes on the truck is $20,000, and the payment is $300/month.
> **Mowers:** He has three new riding mowers. He owes $3,000 on each and he pays $100/month for each one.
> **Other Equipment:** All other equipment is paid for and in good working order, and he is willing to sell all of it to you for $2,000 (and you think it will last 20 mowing months). He estimates it will cost $100/month for maintenance of the truck, mowers, and other equipment.

You have a degree in landscape architecture and design, and you have been the groundskeeper at a golf course for the last five years. You believe with your knowledge and entrepreneurial ideas you could improve greatly on his business model.

In summary, your costs are

- Truck payment: $300/month
- Mowers: $300/month
- Other equipment: $2,000 to purchase now (will last approximately 20 months)
- Chief mower makes $15/hour; staff worker makes $10/hour. Each staff worker can mow 1 lawn/hour. (Average $13/lawn in Scenario 1 or average $25 per lawn in Scenario 2.)

- Maintenance: $100/month
- Insurance: $300/month
- Fuel: $3/gallon; 1 gallon is sufficient to mow one lawn.
- Supplies (trimmer line, etc.): $1.50/lawn
- Advertising: $50/month, but if you do $300/month advertising, you can potentially triple your volume of contracts.
- Owner (you) will spend time doing quality control, marketing, and planning for the future.

Required

To run this business, you'd have to leave your day job, where you earn $4,000/month.

1. Assess whether this mowing business will be profitable if you mow 200 lawns per month.
2. Show a visualization of the change in monthly margin based on a range of labor costs per lawn.

Specify the Question

What is the profit on each lawn if the mowing business mows 200 lawns per month? What is the business's total profit per month?

Obtain the Data

Open Lab 1.1 Data.xlsx and browse its contents.
The data will come from the various assumptions given.

Analyze the Data

We will assess the impact of all of the revenues and costs of each scenario by first estimating the **profit margin** per lawn and then by comparing them in each scenario.

Profit margin: The extent to which revenues exceed expenses, computed as sales minus expenses.

Analysis Task 1: Set Up the Two Scenarios on the Spreadsheet.

1. In cells C4 and D4, insert the number "200" (without the quotation marks), representing the number of lawns you will mow each month.
2. In cells C6 and D6, insert the number "$50", representing the revenue for each lawn.
3. In cell A10, type the word "Truck". In cell B10, insert the number "300" representing the monthly truck payment. To get the cost per lawn, insert the formula "=$B$10/C4" in cell C10, as shown in Lab Exhibit 1.1E.1.

Lab Exhibit 1.1E.1

	A	B	C	D
1	Base Scenario		Scenario 1	Scenario 2
2			**per Lawn**	**per Lawn**
3				
4	Assumption: Total Lawns Serviced Each Month		200	200
5				
6	**Revenue**		$50	$50
7				
8	**Costs**			
9		Monthly		
10	Truck	300	=$B$10/C4	

4. Copy the contents of cell C10 to cell D10.
5. In cell A11, type the word "Mowers". In cell B11, insert the number "300", representing the monthly mowers payment. To get the cost per lawn, insert the formula "=$B$11/C4" in cell C11. Copy the result to cell D11. The resulting spreadsheet should appear as shown in Lab Exhibit 1.1E.2.

Lab Exhibit 1.1E.2

	A	B	C	D
1	Base Scenario		Scenario 1	Scenario 2
2			**Per Lawn**	**Per Lawn**
3				
4	Assumption: Total Lawns Serviced Each Month		200	200
5				
6	**Revenue**		$50	$50
7				
8	**Costs**			
9		Monthly		
10	Truck	300	1.50	1.50
11	Mowers	300	1.50	1.50

Depreciation: The reduction in the value of an asset with the passage of time. Here, we expect the other equipment to last 20 months, so the depreciation equals the $2,000 purchase price divided by 20 months, or $100 per month.

6. In cell A12, type the words "Other Equipment". In cell B12, insert the number "100", representing the monthly **depreciation**. To get the cost per lawn, insert the formula "=$B$12/C4" in cell C12. Copy the result to cell D12.
7. In cell A13, type the words "Labor (Average $13/lawn or $25/lawn)". In cell C13, insert the number "13", representing the labor charge per lawn in Scenario 1. In cell D13, insert the number "25", representing the labor charge per lawn in Scenario 2.
8. In cell A14, type the word "Maintenance". In cell B14, insert the number "100", representing the monthly maintenance expense. To get the cost per lawn, insert the formula "=$B$14/C4" in cell C14. Copy the result to cell D14.
9. In cell A15, type the word "Insurance". In cell B15, insert the number "300", representing the monthly insurance expense. To get the cost per lawn, insert the formula "=$B$15/C4" in cell C15. Copy the result to cell D15. The resulting spreadsheet should appear as shown in Lab Exhibit 1.1E.3.

Lab Exhibit 1.1E.3

	A	B	C	D
1	Base Scenario		Scenario 1	Scenario 2
2			**per Lawn**	**per Lawn**
3				
4	Assumption: Total Lawns Serviced Each Month		200	200
5				
6	**Revenue**		$50	$50
7				
8	**Costs**			
9		Monthly		
10	Truck	300	1.50	1.50
11	Mowers	300	1.50	1.50
12	Other Equipment	100	0.50	0.50
13	Labor (Average $13/lawn or $25/lawn)		13.00	25.00
14	Maintenance	100	0.50	0.50
15	Insurance	300	1.50	1.50
16	Fuel		3.00	3.00
17	Supplies		1.50	1.50

10. In cell A16, type the word "Fuel". In both cells C16 and D16, insert the number "3", representing the fuel cost for each of the two scenarios.
11. In cell A17, type the word "Supplies". In both cells C17 and D17, insert the number "1.50", representing the supplies cost for each of the two scenarios.
12. In cell A18, type the word "Advertising". In cell B18, insert the number "50", representing the monthly advertising payment. To get the advertising cost per lawn, insert the formula "=$B$18/C4" in cell C18. Copy the result to cell D18. The resulting spreadsheet should appear as shown in Lab Exhibit 1.1E.4.

Lab Exhibit 1.1E.4

	A	B	C	D
1	Base Scenario		Scenario 1	Scenario 2
2			**per Lawn**	**per Lawn**
3				
4	Assumption: Total Lawns Serviced Each Month		200	200
5				
6	**Revenue**		$50	$50
7				
8	**Costs**			
9		Monthly		
10	Truck	300	1.50	1.50
11	Mowers	300	1.50	1.50
12	Other Equipment	100	0.50	0.50
13	Labor (Average $13/lawn or $25/lawn)		13.00	25.00
14	Maintenance	100	0.50	0.50
15	Insurance	300	1.50	1.50

13. Insert a single border at the bottom of cells C18 and D18, as shown in Lab Exhibit 1.1E.5. In Excel, go to **Home > Font > Border** and select the bottom border option.

Lab Exhibit 1.1E.5

	A	B	C	D
1	Base Scenario		Scenario 1	Scenario 2
2			**per Lawn**	**per Lawn**
3				
4	Assumption: Total Lawns Serviced Each Month		200	200
5				
6	**Revenue**		$50	$50
7				
8	**Costs**			
9		Monthly		
10	Truck	300	1.50	1.50
11	Mowers	300	1.50	1.50
12	Other Equipment	100	0.50	0.50
13	Labor (Average $13/lawn or $25/lawn)		13.00	25.00
14	Maintenance	100	0.50	0.50
15	Insurance	300	1.50	1.50
16	Fuel		3.00	3.00
17	Supplies		1.50	1.50
18	Advertising	50	0.25	0.25

Analysis Task 2: Calculate Cost and Margin per Lawn and in Total.

1. In cell A19, type the words “Cost per Lawn”. In cell C19, insert the formula “=SUM(C10:C18)” and in cell D19, insert the formula “=SUM(D10:D18)” to calculate the total cost of each scenario.
2. In cell A21, type the words “Margin per Lawn”. In cell C21, insert the formula “=C6-C19” and in cell D19, insert the formula “=D6-D19” to calculate the total margin per lawn for each scenario.
3. In cell A23, type the words “Total Monthly Margin”. In cell C23, insert the formula “=C21*C4” and in cell D19, insert the formula “=D21*D4” to calculate the total monthly margin for each scenario.

The final spreadsheet should appear as shown in Lab Exhibit 1.1E.6.

Lab Exhibit 1.1E.6

	A	B	C	D
1	Base Scenario		Scenario 1	Scenario 2
2			**per Lawn**	**per Lawn**
3				
4	Assumption: Total Lawns Serviced Each Month		200	200
5				
6	**Revenue**		$50	$50
7				
8	**Costs**			
9		Monthly		
10	Truck	300	1.50	1.50
11	Mowers	300	1.50	1.50
12	Other Equipment	100	0.50	0.50
13	Labor (Average $13/lawn or $25/lawn)		13.00	25.00
14	Maintenance	100	0.50	0.50
15	Insurance	300	1.50	1.50
16	Fuel		3.00	3.00
17	Supplies		1.50	1.50
18	Advertising	50	0.25	0.25
19	Cost per Lawn		23.25	35.25
20				
21	Margin per Lawn		26.75	14.75
22				
23	Total Monthly Margin		5350.00	2950.00

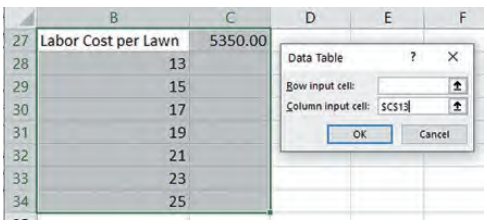
Report the Results

DATA VISUALIZATION

We note that a big difference between the two scenarios is the cost of labor for each lawn. It would be good to know how profitability changes based on the price of labor, which will allow us to assess how profitable the mowing business could be. To conduct this assessment, we will evaluate profitability using Excel's data table for the labor rate from $13 (low end) to $25 (high end) per lawn.

1. Input "Labor Cost per Lawn" in cell B27.
2. Input "13" in cell B28, 15 in cell B29, 17 in cell B30, 19 in cell B31, 21 in cell B32, 23 in cell B33, and 25 in cell B34.
3. In cell C27, insert "=C23". This tells the data table which cell to compute based on the inputs.
4. Highlight cells B27:C34 and select **Data > What-If Analysis > Data Table.**
5. In Column input cell, input "$C$13", as shown in Lab Exhibit 1.1E.7. This tells the data table which input will vary.

Lab Exhibit 1.1E.7

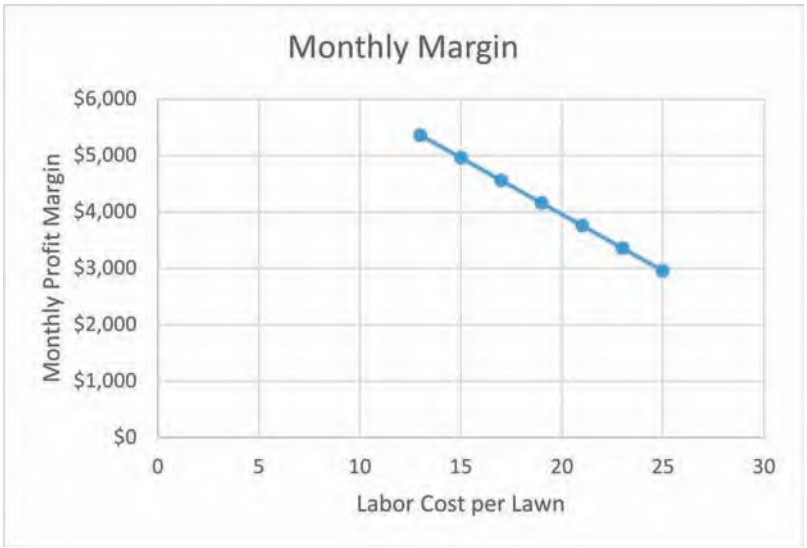


6. Select OK.
7. Copy the cells to another area of the spreadsheet and replace the header of the second column with the words "Monthly Margin" to arrive at the result shown in Lab Exhibit 1.1E.8. Format numbers with commas and a dollar sign ($).

Labor Cost per Lawn	Monthly Margin
13	$5,350
15	$4,950
17	$4,550
19	$4,150
21	$3,750
23	$3,350
25	$2,950

Lab Exhibit 1.1E.8

8. Select and highlight cells B28:C34, then select **Insert > Charts > Scatterplot** (with smooth lines and markers), and click OK. Label the horizontal axis as "Labor Cost per Lawn" to arrive at the visualization shown in Lab Exhibit 1.1E.9. Label the vertical axis as "Monthly Profit Margin". You might use this visualization in a brief written report or as part of a PowerPoint presentation.



Lab Exhibit 1.1E.9

Assessment

1. Take a screenshot of the completed spreadsheet, paste it into a Word document named "Lab 1.1 Excel Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the scatterplot showing labor cost per lawn and monthly margin, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 1.1 Submission.docx via Connect, if it is assigned.

Alternate Lab 1.1 Excel

Assessing Profitability of a Mowing Business

Apply the same steps in Lab 1.1 to the Lab 1.1 Data.xlsx data set with two modifications. The same scenarios exist as before, with labor costing either $13/lawn or $25/lawn.

Assume that your mowing business will mow 350 lawns per month rather than 200 lawns per month. Also assume that the cost for advertising increases from $50 to $300 per month. Label these two new scenarios Scenario 3 (with labor costing $13/lawn) and Scenario 4 (with labor costing $25/lawn).

Specify the Question

What is the profit on each lawn if the mowing business mows 350 lawns per month? What is the business's total profit per month?

Required

1. Assess whether this mowing business will be profitable if you mow 350 lawns per month.
2. Show a visualization of the change in monthly margin based on a range of labor costs per lawn.

Assessment

1. Take a screenshot of the completed spreadsheet, paste it into a Word document named "Alt Lab 1.1 Excel Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the scatterplot showing labor cost per lawn and monthly margin, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 1.1 Submission.docx via Connect, if it is assigned.

Lab 1.2 Excel

Assessing Customer Profitability

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Customer Profitability, Inventory Management, Customer Relationship Management

Decision-Making Context

Merchandising companies sell products to customers. They make money by selling the products for more than what they paid. In this lab, we compute the **gross margin** for each sale and then identify the company's most profitable customers. This analysis allows merchandising companies to focus their efforts on selling more products to those customers, which in turn helps them make more profit overall.

Gross margin: Net sales minus the cost of goods sold. Gross margin is the amount of money a company retains after deducting the direct costs of producing a good or service.

Fancy Fruits is a (fictitious) mail-order business that sells exotic fruit from around the world to customers throughout the United States. The company keeps a sales journal that records each sale and its related sales and cost details.

Required

1. Calculate gross margin by customer to identify the most profitable customer.
2. Create a data visualization to report the results to management.

Specify the Question

Which customers are the most profitable?

Obtain the Data

This sales journal is for March 1–4, 2025. The sales journal provides a list of all individual sales (in addition to other data) for the four-day period.

Open Excel file Lab 1.2 Data.xlsx. The data dictionary follows. Lab Exhibit 1.2E.1 shows a snippet of the spreadsheet.

Data Dictionary

Invoice #:	Invoice number for each sale of fruit
Customer #:	Customer number making purchase
Zip Code:	Zip code of customer making purchase
Date of Sale:	Date of **invoice**
SKU:	Fancy Fruits' SKU (SKU = stock-keeping unit, a unique code for each fruit product sold)
Description:	Description of each fruit sold
Quantity (Pounds):	Weight of each product sold
Sales:	Total sales price of product sold
Cost:	Total cost of product sold

Invoice: A list of items that were purchased (or sold) along with the amount due from the customer; also called a **bill**.

Lab Exhibit 1.2E.1

	A	B	C	D	E	F	G	H	I
1	Invoice #	Customer #	Zip Code	Date of Sale	SKU	Description	Quantity (Pounds)	Sales	Cost
2	3021	2001	67510	3/1/2025	FRT-APL	Apple	15	25.52	16.36
3	3021	2001	67510	3/1/2025	FRT-GOR	Goji berry	5	5.92	3.71
4	3021	2001	67510	3/1/2025	FRT-OLV	Olive	35	53.92	33.26
5	3021	2001	67510	3/1/2025	FRT-REN	Redcurrant	5	7.12	4.26
6	3022	2091	67621	3/1/2025	FRT-FII	Fig	35	84.37	54.26
7	3023	2025	67511	3/1/2025	FRT-JUR	Juniper berry	50	65.02	43.01
8	3023	2025	67511	3/1/2025	FRT-PAY	Papaya	35	37.12	22.06
9	3023	2025	67511	3/1/2025	FRT-BAN	Banana	5	2.52	1.56
10	3023	2025	67511	3/1/2025	FRT-PEO	Persimmon	35	69.67	43.06
11	3024	2079	66401	3/1/2025	FRT-SAM	Satsuma	50	34.52	22.01
12	3024	2079	66401	3/1/2025	FRT-MII	Miracle fruit	15	30.77	18.01
13	3024	2079	66401	3/1/2025	FRT-CHI	Chico fruit	15	26.27	15.46
14	3024	2079	66401	3/1/2025	FRT-DUA	Durian	25	54.27	33.01
15	3024	2079	66401	3/1/2025	FRT-BAN	Banana	35	17.52	10.86
16	3025	2099	66710	3/1/2025	FRT-SAM	Satsuma	30	20.72	13.21
17	3025	2099	66710	3/1/2025	FRT-ELR	Elderberry	25	60.27	38.01
18	3025	2099	66710	3/1/2025	FRT-NAC	Nance	5	5.92	3.61
19	3025	2099	66710	3/1/2025	FRT-BAN	Banana	45	22.52	13.96

Analyze the Data

Analysis Task 1: Compute the Gross Margin of Each Sale.

We first need to compute the gross margin of each sale.

1. Start by going to cell J1. Insert the title "Gross Margin" and insert a bottom border. In cell J2, we'll compute the gross margin on each line item sold. To do so, we must subtract the cost in cell I2 from the sales in cell H2. To do so, insert the formula "=H2-I2" (as shown in Lab Exhibit 1.2E.2) and click Enter.

Lab Exhibit 1.2E.2

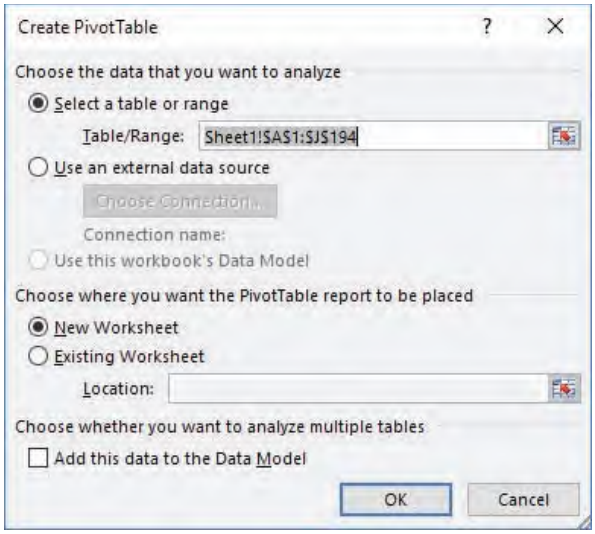
	H	I	J	K
1	Sales	Cost	Gross Margin	
2	25.52	16.36	=H2-I2	
3	5.92	3.71		
4	53.92	33.26		
5	7.12	4.26		
6	84.37	54.26		

2. Copy that formula down for every sale from cell J2 to cells J3:J194.
3. Highlight column J by clicking on column J. Right-click and select "Format Cells". Select the "Number" tab. Then select "Number" and set Decimal places to "2". Click OK.

Analysis Task 2: Calculate Profitability by Customer.

To know the profitability of each customer, we need to compute the gross margin for each customer. To do so, we will use a pivot table.

1. In the menu, select **Insert > Pivot Table.** A dialog box will open as shown in Lab Exhibit 1.2E.3.



Lab Exhibit 1.2E.3

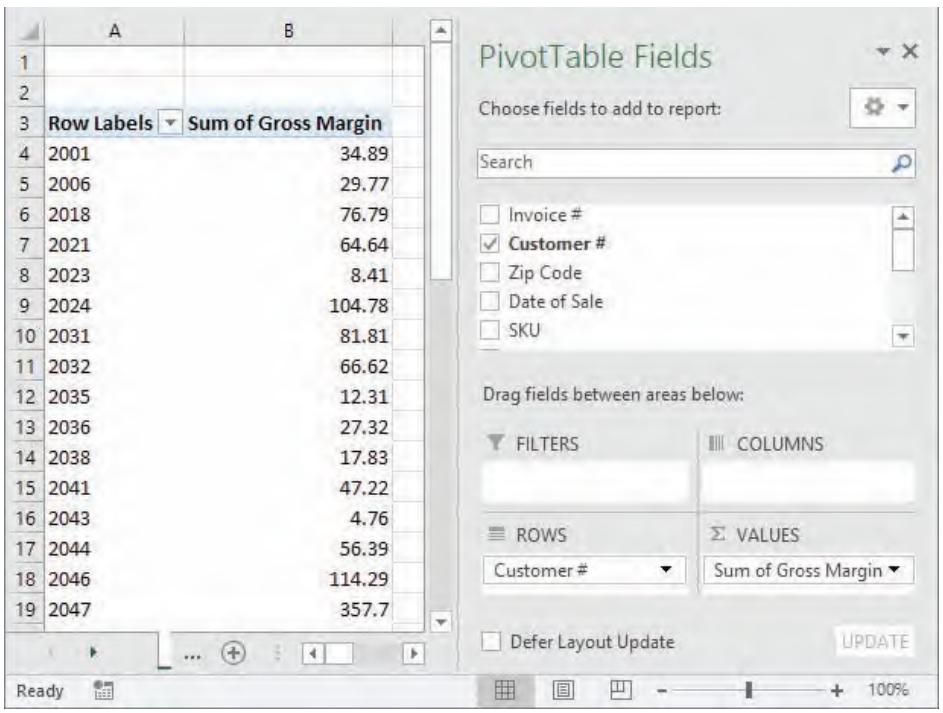
2. In the dialog box, select the Table/Range that includes the account titles and all of the data ($A$1:$J$194). Select "New Worksheet", and then click OK. The empty pivot table will open up in a new worksheet, ready for the next step (as shown in Lab Exhibit 1.2E.4).

Lab Exhibit 1.2E.4

Inputs into Pivot Table
Columns:
Rows: [Customer #]
ΣValues: [Gross Margin]

3. Drag [Customer #] from **FIELD NAME** into the **Rows** and [Gross Margin] from **FIELD NAME** into ΣValues fields in the pivot table. The ΣValues will default to "Sum of Gross Margin". The resulting pivot table will look like Lab Exhibit 1.2E.5.
4. (Optional Step) Highlight cell A3 and insert the title "Customer Number".

Lab Exhibit 1.2E.5



Analysis Task 3: Sort Total Gross Margin by Customer.

1. Our last analysis task is to sort Gross Margin from highest to lowest to find the most profitable customer. To do so, select one of the numbers in column B, and then select **Data > Sort.** To get the most profitable customers on top, select descending.
2. Select the numbers in column B. Right-click and select "Format Cells". Select the "Number" tab. Then select "Number" and set Decimal places to "2". Click OK.
3. The pivot table will now appear as shown in Lab Exhibit 1.2E.6.

Lab Exhibit 1.2E.6

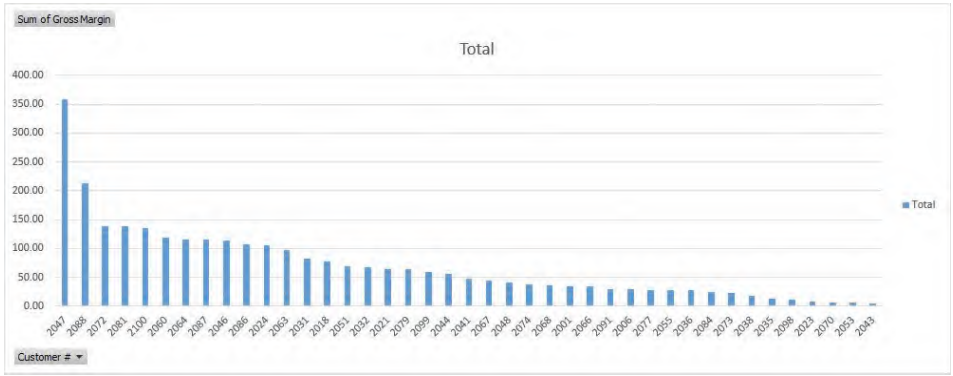
	A	B
1		
2		
3	Row Labels	Sum of Gross Margin
4	2047	357.70
5	2088	212.23
6	2072	139.01
7	2081	137.78
8	2100	135.47
9	2060	118.74
10	2064	115.33
11	2087	114.85
12	2046	114.29
13	2086	107.41
14	2024	104.78
15	2063	96.45
16	2031	81.81
17	2018	76.79
18	2051	69.83
19	2032	66.62
20	2021	64.64
21	2079	64.00
22	2099	58.56
23	2044	56.39
24	2041	47.22
25	2067	44.94
26	2048	41.12
27	2074	38.18
28	2068	36.62

Report the Results

To report the results to management, we're going to create a visualization using a pivot chart. Follow these steps:

1. Select a cell in the pivot table.
2. Select **PivotTable Analyze > Tools > PivotChart.**
3. Select Clustered Column.
4. Click OK.

The resulting pivot chart will be as shown in Lab Exhibit 1.2E.7.



Lab Exhibit 1.2E.7
Profitability by Customer ID

We can report the results of Fancy Fruits' most profitable customers to the sales department and to company management. To keep these customers engaged and satisfied, we could send additional advertising emails to them, enhance customer service rewards, or give them discounts on future purchases.

Assessment

1. Take a screenshot of your pivot table (using a snipping tool or other software), paste it into a Word document named "Lab 1.2 Excel Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of your pivot chart (using a snipping tool or other software), paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 1.2 Excel Submission.docx via Connect, if it is assigned.

Lab 1.2 Tableau

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Assessing Customer Profitability

Keywords

Customer Profitability, Inventory Management, Customer Relationship Management

Decision-Making Context

Merchandising companies sell products to customers. They make money by selling the products for more than what they paid. In this lab, we compute the gross margin for each sale and then identify the company's most profitable customers. This analysis allows merchandising companies to focus their efforts on selling more products to those customers, which in turn helps them make more profits.

Required

Gross margin: Net sales minus the cost of goods sold. Gross margin is the amount of money a company retains after deducting the direct costs of producing a good or service.

1. Create a calculated field for the **gross margin**.
2. Create a summary table that shows the gross margin per customer.
3. Create a horizontal bar graph that sorts the gross margin by customer.

Specify the Question

Which customers are the most profitable?

Obtain the Data

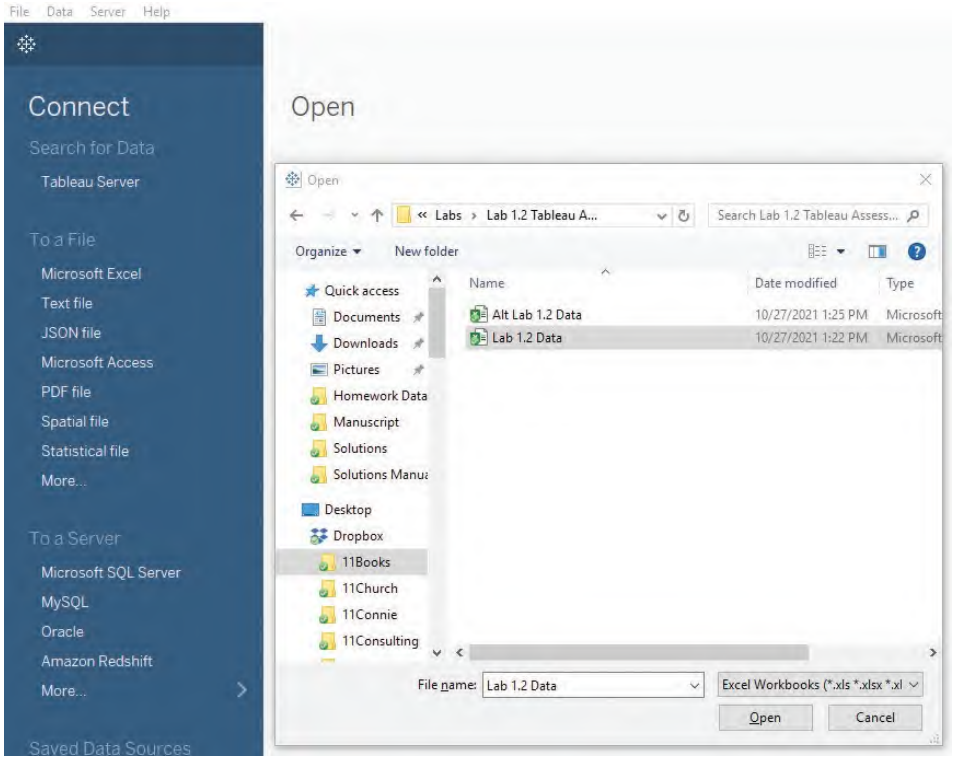
Fancy Fruits is a (fictitious) mail-order business that sells exotic fruit from around the world to customers throughout the United States. The company keeps a sales journal that records each sale and its related sales and cost details. This sales journal is for March 1–4, 2025.

Data Dictionary

Invoice #:	Invoice number for each sale of fruit
Customer #:	Customer number making purchase
Zip Code:	Zip code of customer making purchase
Date of Sale:	Date of **invoice**
SKU:	Fancy Fruits' SKU (SKU = stock-keeping unit, a unique code for each fruit product sold)
Description:	Description of each fruit sold
Quantity (Pounds):	Weight of each product sold
Sales:	Total sales price of product sold
Cost:	Total cost of product sold

Invoice: A list of items that were purchased (or sold) along with the amount due from the customer; also called a **bill**.

1. Open Tableau. From Tableau, select **Connect > Microsoft Excel,** then browse to the location where you have saved file Lab 1.2 Data.xlsx, and click Open (as shown in Lab Exhibit 1.2T.1).



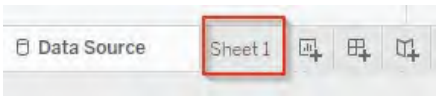
Lab Exhibit 1.2T.1

Tableau Software

Analyze the Data

Analysis Task 1: Compute the Gross Margin of Each Sale.

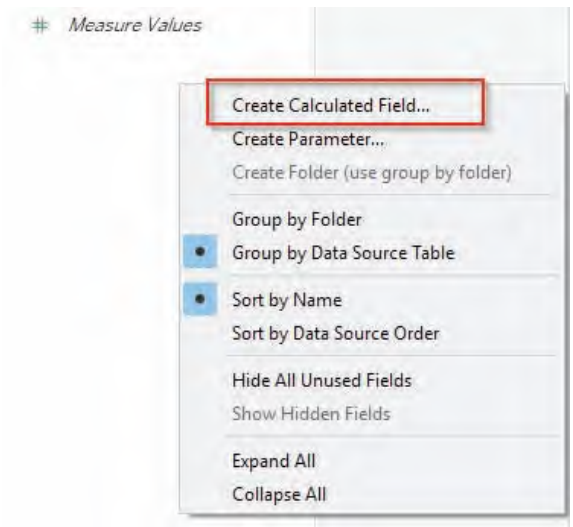
1. Click Sheet 1 to begin working with the data (Lab Exhibit 1.2T.2).



Lab Exhibit 1.2T.2

Tableau Software

2. Right-click below the listed variables and select "Create Calculated Field" to create Calculated Field (Lab Exhibit 1.2T.3).



Lab Exhibit 1.2T.3

Tableau Software

Name your new calculated field "**Gross Margin**" and hit Enter.

Gross Margin is calculated by subtracting the **Cost** field from the **Sales** field. Type in "**Sales – Cost"** and click "OK" (Lab Exhibit 1.2T.4).

Tableau will automatically configure the words Cost and Sales into references to the fields in the data set. The sentence "The calculation is valid" indicates that the formula is working. Click OK.

Lab Exhibit 1.2T.4

Tableau Software

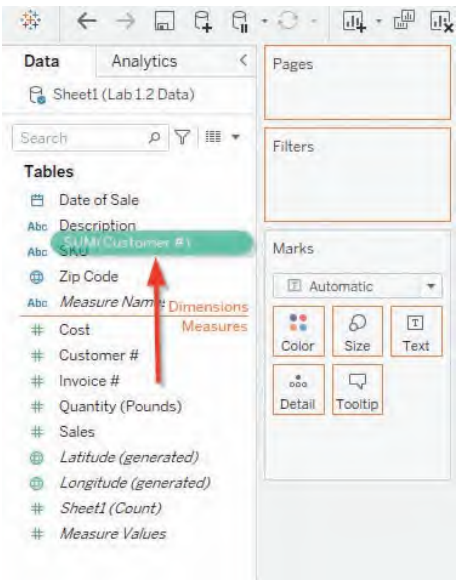


Analysis Task 2: Calculate Profitability by Customer.

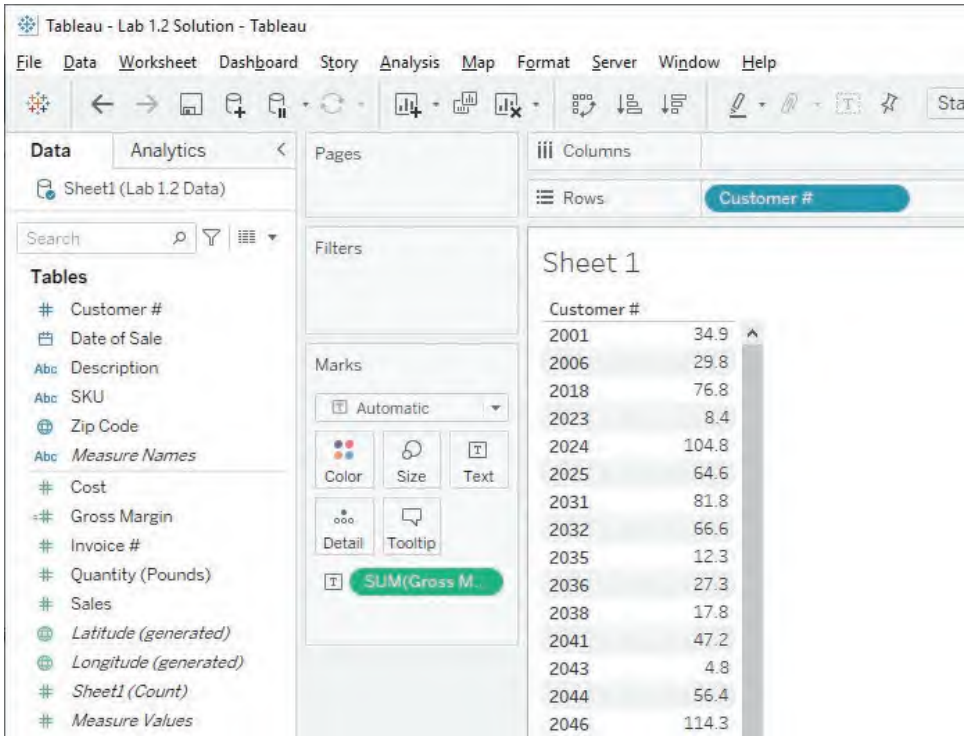
1. Now that we have created our calculated field, we can build a table to show profitability by customer.
2. If **Customer** has defaulted to the **Measures**, you will need to drag and drop it into the **Dimensions** section of Tableau (Lab Exhibit 1.2T.5).

Lab Exhibit 1.2T.5

Tableau Software



Double-click **Customer** to see a Customer list, then double-click **Gross Margin** to see the corresponding gross margin value for each customer. It will appear as a "pill" on the marks shelf (Lab Exhibit 1.2T.6).

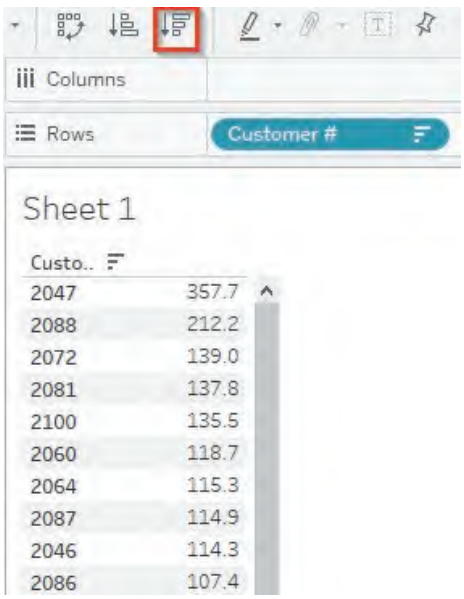


Lab Exhibit 1.2T.6

Tableau Software

Analysis Task 3: Sort Total Gross Margin by Customer.

1. Our last step is to sort Gross Margin from highest to lowest to find the most profitable customer. To do so, select the Sort icon (Lab Exhibit 1.2T.7).



Lab Exhibit 1.2T.7

Tableau Software

Report the Results

To report the results to management, we're going to create a visualization using horizontal bars. To do so, go to the "Show Me" icon in the top right and select the icon for horizontal bars (Lab Exhibit 1.2T.8). The resulting visualization will be as shown in Lab Exhibit 1.2T.9.

DATA VISUALIZATION

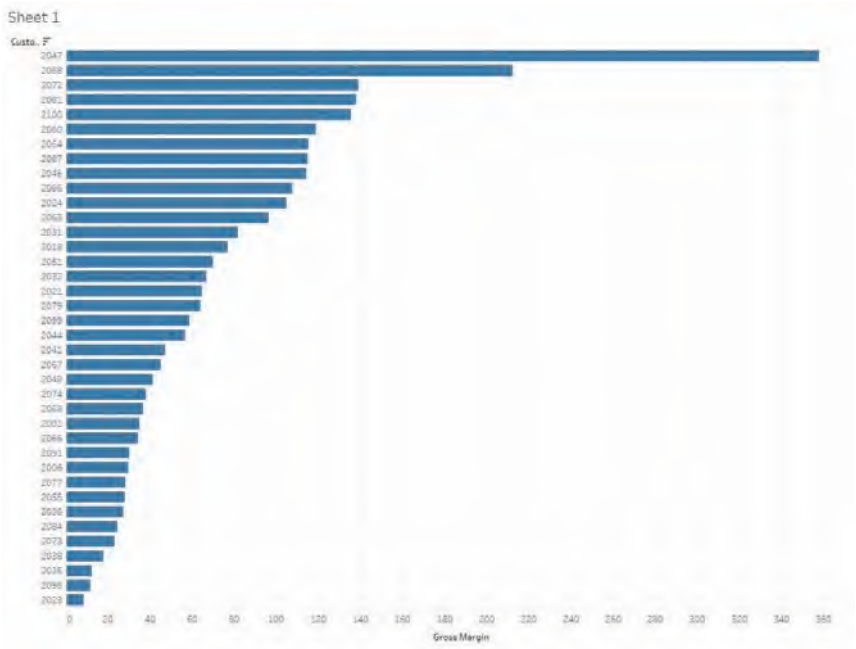
Lab Exhibit 1.2T.8

Tableau Software



Lab Exhibit 1.2T.9

Tableau Software



We have now identified the most profitable customers and can share this list with management. We may want to promote additional products to these customers.

Assessment

1. Take a screenshot of your customer profitability table, paste it into a Word document named "Lab 1.2 Tableau Submission.docx", and label the screenshot Submission 1.
2. Create a horizontal bar graph that sorts the gross margin by customer, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 1.2 Tableau Submission.docx via Connect, if it is assigned.

Lab 1.2 Power BI

Assessing Customer Profitability

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Customer Profitability, Inventory Management, Customer Relationship Management

Decision-Making Context

Merchandising companies sell products to customers. They make money by selling the products for more than what they paid. In this lab, we compute the gross margin for each sale and then identify the company's most profitable customers. This analysis allows merchandising companies to focus their efforts on selling more products to those customers, which in turn helps them make more profits.

Required

1. Create a calculated field for the **gross margin**.
2. Create a summary table that shows the gross margin per customer.
3. Create a horizontal bar graph that sorts the gross margin by customer.

Gross margin: Net sales minus the cost of goods sold. Gross margin is the amount of money a company retains after deducting the direct costs of producing a good or service.

Specify the Question

Which customers are the most profitable?

Obtain the Data

Fancy Fruits is a (fictitious) mail-order business that sells exotic fruit from around the world to customers throughout the United States. The company keeps a sales journal that records each sale and its related sales and cost details. This sales journal is for March 1–4, 2025.

Data Dictionary

Invoice #:	Invoice number for each sale of fruit
Customer #:	Customer number making purchase
Zip Code:	Zip code of customer making purchase
Date of Sale:	Date of **invoice**
SKU:	Fancy Fruits' SKU (SKU = stock-keeping unit, a unique code for each fruit product sold)
Description:	Description of each fruit sold
Quantity (Pounds):	Weight of each product sold
Sales:	Total sales price of product sold
Cost:	Total cost of product sold

Invoice: A list of items that were purchased (or sold) along with the amount due from the customer; also called a **bill**.

1. Open Power BI Desktop. Connect Power BI to the Excel file. From Power BI Desktop, select **Home > Get Data > Excel Workbook**, then browse to the location where you have saved file Lab 1.2 Data.xlsx, and click Open.
2. In the Navigator window, check the box next to Sheet 1 and click Load.

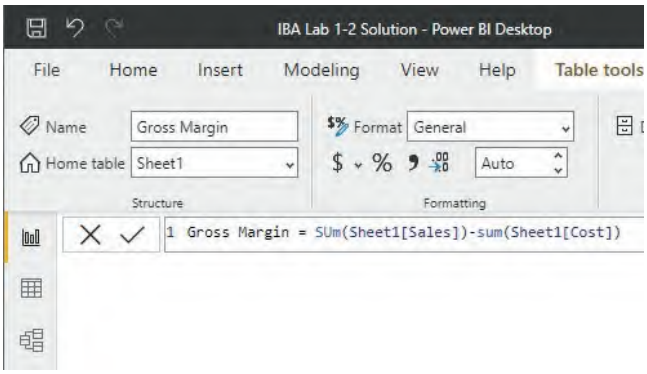
Analyze the Data

Analysis Task 1: Compute the Gross Margin of Each Sale.

1. From the Ribbon, click **Home > Calculations > New Measure** to create Calculated Field (Lab Exhibit 1.2P.1).
2. Name your new calculated field **Gross Margin.**
3. Gross Margin is calculated by subtracting the **Cost** field from the **Sales** field. Type in "Gross Margin = SUM(Sheet1[Sales]) – SUM(Sheet1[Cost])" and press Enter.

Lab Exhibit 1.2P.1

Microsoft Power BI



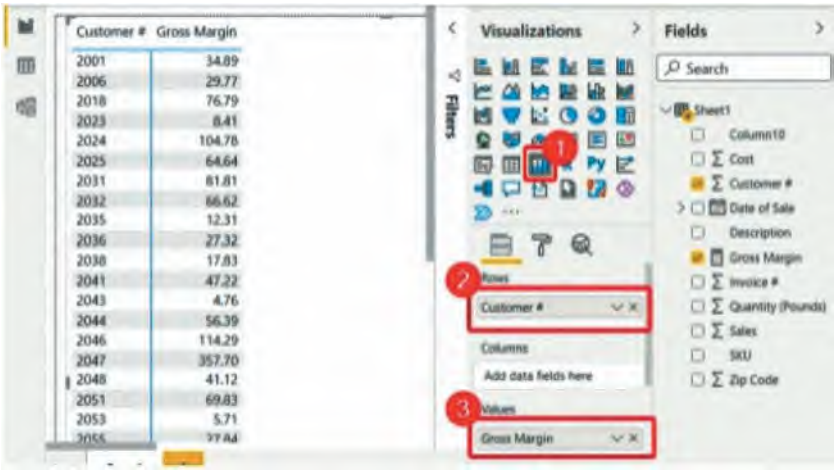
Analysis Task 2: Calculate Profitability by Customer.

Now that we have created our calculated field, we can build a table to show us profitability by customer.

1. Click in the blank area of your page, and then click the Matrix button in the visualizations pane to add it to your page. Resize the table to fill the left side of your page.
2. Drag the Customer # from the field list to the Rows box in the Visualizations pane to see a Customer list.
3. Drag Gross Margin from the field list to the Values box in the Visualizations pane to see the corresponding gross margin value for each customer (Lab Exhibit 1.2P.2).

Lab Exhibit 1.2P.2

Microsoft Power BI



Analysis Task 3: Sort Total Gross Margin by Customer.

Our last step is to sort Gross Margin from highest to lowest to find the most profitable customer. To do so, click the Gross Margin column header in the table once.

Report the Results


DATA VISUALIZATION

To report the results to management, we're going to create a visualization using horizontal bars. To do so, right-click the matrix that you created and select **Copy > Copy Visual**, then click Paste to add it to your dashboard. From the visualization pane, select Stacked Column Chart.

We have now identified the most profitable customers and can share this list with management. We may want to promote additional products to these customers.

Assessment

1. Take a screenshot of your dashboard, paste it into a Word document named "Lab 1.2 Power BI Submission.docx", and label the screenshot Submission 1.
2. Create a horizontal bar graph that sorts the gross margin by customer, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 1.2 Power BI Submission.docx via Connect, if it is assigned.

Alternate Lab 1.2

Assessing Customer Profitability

Apply the same steps from Lab 1.2 (Excel, Tableau, or Power BI) to the Alt Lab 1.2 Data.xlsx data set using Excel, Tableau, or Power BI.

Open Excel File Alt Lab 1.2 Data.xlsx and follow similar directions to those in Lab 1.2 above using Excel, Tableau, or Power BI. ThrustMaster is a fictitious company that sells game controllers, joysticks, and steering wheels for PCs and for Xbox, Nintendo, and PlayStation consoles. A summary of the company's sales is included in the data set.

Required

1. Calculate **gross margin** by customer to identify the most profitable customer.

Gross margin: Net sales minus the cost of goods sold. Gross margin is the amount of money a company retains after deducting the direct costs of producing a good or service.

Assessment

1. Take a screenshot of your chart, paste it into a Word document named "Alt Lab 1.2 Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 1.2 Submission.docx via Connect, if it is assigned.

Lab 1.3 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Gross margin: Net sales minus the cost of goods sold. Gross margin is the amount of money a company retains after deducting the direct costs of producing a good or service.

SKU: Stock-keeping unit; a unique code for each product sold.

Assessing Product (SKU) Profitability

Keywords

SKU, Product Profitability, Inventory Management

Decision-Making Context

Merchandising companies sell products to customers. They make money by selling the products for more than what they paid. In this lab, we compute the **gross margin** for each product to identify the company's most profitable products. This analysis allows merchandising companies to focus their efforts on selling the most profitable products, which in turn helps them make more profits.

Required

1. Create a histogram of the gross margin percentages using bins of 0.025 size.
2. Create a list of **SKUs** in the highest 0.025 bin (from 0.40 to 0.425).
3. Create a visualization of the histogram.

Specify the Question

Which products (SKUs) are the most profitable to sell?

Obtain the Data

Fancy Fruits is a fictitious mail-order business that sells exotic fruit from around the world to customers throughout the United States. The company keeps a price list for each product. We will work from this price list to find the items that are most profitable per dollar of sales. This is the data dictionary:

Data Dictionary

SKU:	Fancy Fruits' SKU (unique code for each fruit product sold)
Description:	Description of each fruit sold
Quantity (Pounds):	Weight of each product sold
Sales Price per Pound:	Total sales price per pound of product
Cost per Pound:	Total cost of product sold per pound of product

Open Excel file Lab 1.3 Data.xlsx, which includes Fancy Fruits, price list. A snippet of the spreadsheet appears in Lab Exhibit 1.3E.1.

Lab Exhibit 1.3E.1

	A	B	C	D
1	SKU	Description	Sales Price per Pou	Cost per Pound
2	FRT-APL	Apple	1.7	1.09
3	FRT-APO	Apricot	2.8	1.76
4	FRT-AVD	Avocado	1	0.59
5	FRT-BAN	Banana	0.5	0.31
6	FRT-BIT	Bilberty	2	1.17
7	FRT-BLR	Blackberry	2	1.29
8	FRT-BLN	Blackcurrent	2	1.18
9	FRT-BLB	Blueberry	1.75	1.07
10	FRT-BOR	Boysenberry	1	0.62
11	FRT-CUN	Currant	1.75	1.04
12	FRT-CHR	Cherry	2.5	1.56
13	FRT-CHY	Cherimoya	1.75	1.11
14	FRT-CHI	Chico fruit	1.75	1.03
15	FRT-CLR	Cloudberry	1.75	1.09
16	FRT-COU	Coconut	2.41	1.54
17	FRT-CRR	Cranberry	1.01	0.61
18	FRT-CUE	Cucumber	1.82	1.09
19	FRT-SUL	Sugar apple	1.01	0.62
20	FRT-DAO	Damson	2.04	1.34
21	FRT-DAT	Date	0.83	0.54
22	FRT-DRI	Dragonfruit	1.63	0.98

Analyze the Data

Analysis Task 1: Compute the Gross Margin and Gross Margin Percentage for Each Product (SKU).

1. We first need to compute the gross margin for each SKU. A SKU, or stock-keeping unit, is an alphanumeric identification that tracks the inventory of a particular product.
2. Go to cell E1 and insert the title "Gross Margin". Also, insert a bottom border. In cell E2, we'll compute the gross margin on each line item sold. To do so, we must subtract the cost in cell D2 from the sales in cell C2. Insert the formula "=C2-D2" and hit Enter (Lab Exhibit 1.3E.2). Next, copy that formula down for every sale below from cell E3:E82.

Lab Exhibit 1.3E.2

	A	B	C	D	E
1	SKU	Description	Sales Price per Pou	Cost per Pound	Gross Margin
2	FRT-APL	Apple	1.7	1.09	=C2-D2
3	FRT-APO	Apricot	2.8	1.76	1.04
4	FRT-AVD	Avocado	1	0.59	0.41
5	FRT-BAN	Banana	0.5	0.31	0.19
6	FRT-BIT	Bilberty	2	1.17	0.83
7	FRT-BLR	Blackberry	2	1.29	0.71
8	FRT-BLN	Blackcurrent	2	1.18	0.82

3. We also need to compute the gross margin percentage, which is the percent of every dollar of sales that is gross margin. To compute the gross margin percentage, go to cell F2 and enter "=E2/C2", as shown in Lab Exhibit 1.3E.3.

Lab Exhibit 1.3E.3

	A	B	C	D	E	F
1	SKU	Description	Sales Price per Pou	Cost per Pound	Gross Margin	Gross Margin Percentage
2	FRT-APL	Apple	1.7	1.09	0.61	=E2/C2
3	FRT-APO	Apricot	2.8	1.76	1.04	0.371
4	FRT-AVD	Avocado	1	0.59	0.41	0.410
5	FRT-BAN	Banana	0.5	0.31	0.19	0.380

4. Then copy that formula down for every sale from cell F3:F82.

Analysis Task 2: Sort the Gross Margin Percentages into Buckets Using Excel's Data Analysis ToolPak.

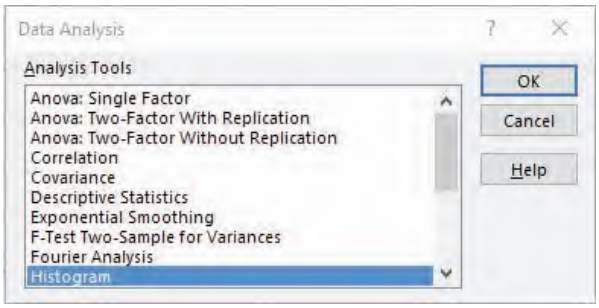
1. To report the results, it is important to be able to visualize the results. To do so, we will plot the gross margin percentages using a histogram to see the frequency of gross margin percentages in each bin.
2. Insert in cell I1 the word "Bins". Also, insert a bottom border. Insert the following bins beginning in cell I2 and continuing through cell I6: 0, 0.35, 0.375, 0.4, and 0.425, as shown in Lab Exhibit 1.3E.4.

Lab Exhibit 1.3E.4

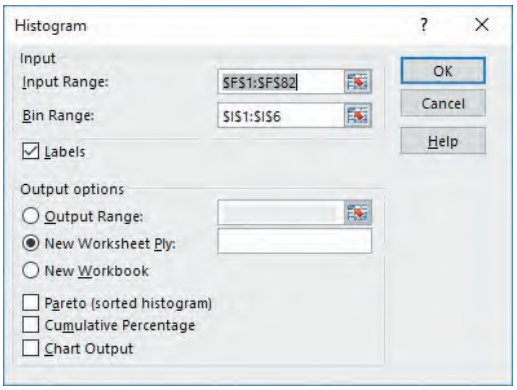
	I
1	Bins
2	0
3	0.35
4	0.375
5	0.4
6	0.425

3. We will need the Data Analysis ToolPak to use the Histogram tool. If you haven't installed the Data Analysis ToolPak, Appendix E at the end of this textbook provides instructions for installing the ToolPak. Once the Data Analysis ToolPak is installed, select **Data > Analysis > Data Analysis > Histogram,** as shown in Lab Exhibit 1.3E.5.

Lab Exhibit 1.3E.5



4. For Input Range, input "$F$1:$F$82" to get all of the gross margin percentage possibilities.
5. For Bin Range, input "$I$1:$I$6" to get all of the bin possibilities.
6. Click "Labels", as we have included the labels in our bin range. Steps 4–6 are shown in Lab Exhibit 1.3E.6.



Lab Exhibit 1.3E.6

7. Click OK to get the result that is shown in Lab Exhibit 1.3E.7.

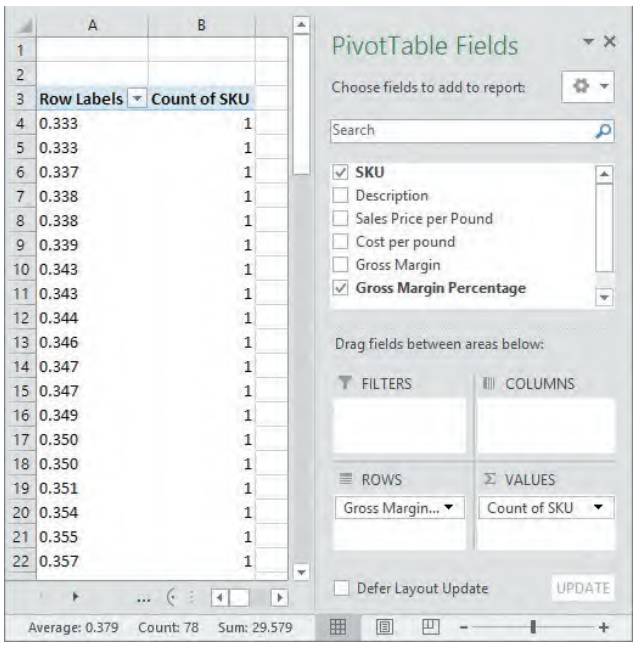
Lab Exhibit 1.3E.7

	A	B
1	*Bins*	*Frequency*
2	0	0
3	0.35	13
4	0.375	21
5	0.4	24
6	0.425	23
7	More	0

Analysis Task 3: Sort the Gross Margin Percentages into Buckets Using a Pivot Table.

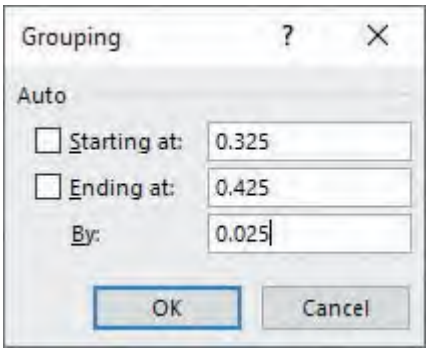
1. We'll now work on the second requirement: Get a list of SKUs in the highest 0.025 bin (from 0.40 to 0.425). We could sort the gross margin percentages from lowest to highest and copy the 24 SKUs in that range. A second, more sophisticated way to get that list is to use a pivot table.
2. To do so, first click in the table itself first and then click on **Insert > Pivot Table** after clicking on one cell of the data.
3. Drag [Gross Margin Percentage] from **FIELD NAME** into the **Rows** and [SKU] from **FIELD NAME** into ΣValues fields in the pivot table. The ΣValues will default to "Count of SKU". The resulting pivot table will look like Lab Exhibit 1.3E.8.

Lab Exhibit 1.3E.8



4. Right-click on one of the numbers under Row Labels in column A, and select Group.
5. Insert 0.325 for Starting at:, 0.425 for Ending at:, and 0.025 for By in the Grouping dialog box, as shown in Lab Exhibit 1.3E.9.

Lab Exhibit 1.3E.9



6. Click OK to get the result shown in Lab Exhibit 1.3E.10.

Lab Exhibit 1.3E.10

Row Labels	Count of SKU
0.325-0.35	13
0.35-0.375	21
0.375-0.4	24
0.4-0.425	23
Grand Total	**81**

Note that the counts of SKUs in each bin are the same as those shown in the histogram in Lab Exhibit 1.3E.12.

7. The second requirement asks us to get a detailed listing of the products that are in the 0.40 to 0.425 gross margin percentage range. To do so, all we need to do is double-click on cell B7 with the 23 under "Count of SKU". Double-clicking on it will give us the list of the products in the 0.4 to 0.425 range, as shown in Lab Exhibit 1.3E.11.

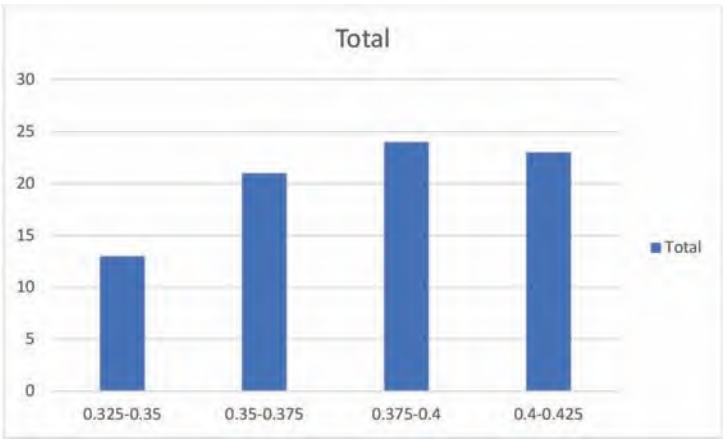
Lab Exhibit 1.3E.11

SKU	Description	Sales Price per Pound	Cost per Pound	Gross Margin	Gross Margin Percentage
FRT-SAR	Salal berry	1.12	0.66	0.46	0.411
FRT-REN	Redcurrant	1.42	0.85	0.57	0.401
FRT-AVD	Avocado	1	0.59	0.41	0.410
FRT-PUE	Purple mangos	1.01	0.59	0.42	0.416
FRT-BIT	Bilberty	2	1.17	0.83	0.415
FRT-PLT	Plumcot (or Pl	2.1	1.24	0.86	0.410
FRT-BLN	Blackcurrent	2	1.18	0.82	0.410
FRT-PIL	Pineapple	0.77	0.46	0.31	0.403
FRT-PLM	PlumPrune (dr	1.31	0.77	0.54	0.412
FRT-CUN	Currant	1.75	1.04	0.71	0.406
FRT-PEA	Pear	1.83	1.07	0.76	0.415
FRT-PAY	Papaya	1.06	0.63	0.43	0.406
FRT-CHI	Chico fruit	1.75	1.03	0.72	0.411
FRT-MII	Miracle fruit	2.05	1.2	0.85	0.415
FRT-LOG	Longan	1.85	1.1	0.75	0.405
FRT-LIM	Lime	0.96	0.57	0.39	0.406
FRT-CUE	Cucumber	1.82	1.09	0.73	0.401
FRT-KII	Kiwifruit	2.32	1.37	0.95	0.409
FRT-JAU	Jambul	1.95	1.13	0.82	0.421
FRT-JAI	Jackfruit	0.63	0.37	0.26	0.413
FRT-JAB	Jabuticaba	0.89	0.52	0.37	0.416
FRT-GUV	Guava	0.94	0.56	0.38	0.404
FRT-GRI	GrapeRaisin	1.25	0.74	0.51	0.408

Report the Results: Visualizing the Histogram Using a Pivot Chart

It is important to visualize the counts of products in each percentage bin. To visualize, highlight any cell in the pivot table, click Pivot chart, and select Clustered column to see the visualization shown in Lab Exhibit 1.3E.12.

DATA VISUALIZATION



Lab Exhibit 1.3E.12

The histogram gives us a good illustration of the distribution of the gross margin percentages.

We report the results to the sales staff and company management to identify the products (SKUs) that are most profitable. The company may want to emphasize the marketing of these products in the future.

Assessment

1. Take a screenshot of your detailed listing of fruit products sold from the 0.40 to 0.425 gross margin percentage range, paste it into a Word document named "Lab 1.3 Excel Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of your histogram as a pivot chart, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 1.3 Excel Submission.docx via Connect, if it is assigned.

Lab 1.3 Tableau

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Gross margin: Net sales minus the cost of goods sold. Gross margin is the amount of money a company retains after deducting the direct costs of producing a good or service.

SKU: Stock-keeping unit; a unique code for each product sold.

Assessing Product (SKU) Profitability

Keywords

Customer Profitability, Inventory Management, Customer Relationship Management

Decision-Making Context

Merchandising companies sell products to customers. They make money by selling the products for more than what they paid. In this lab, we compute the **gross margin** for each product to identify the company's most profitable products. This analysis allows merchandising companies to focus their efforts on selling the most profitable products, which in turn helps them make more profits.

Required

1. Create calculated fields for the **gross margin** and gross margin percentage.
2. Create a histogram to visualize the distribution of gross margin percentage across **SKUs**.
3. Adjust the automated bin size to 0.025.
4. Create a list of SKUs in the highest bin.

Specify the Question

Which products (SKUs) are the most profitable to sell?

Obtain the Data

Fancy Fruits is a fictitious mail-order business that sells exotic fruit from around the world to customers throughout the United States. The company keeps a price list for each product. We will work from this price list to find the items that are most profitable per dollar of sales. This is the data dictionary:

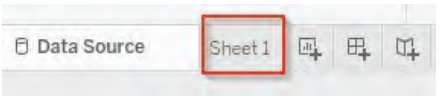
Data Dictionary

SKU:	Fancy Fruits' SKU (unique code for each fruit product sold)
Description:	Description of each fruit sold
Quantity (Pounds):	Weight of each product sold
Sales Price per Pound:	Total sales price per pound of product
Cost per Pound:	Total cost of product sold per pound of product

1. Open Tableau. Then open Lab 1.3 Data.xlsx.
2. Click Sheet 1 to begin working with the data (Lab Exhibit 1.3T.1).

Lab Exhibit 1.3T.1

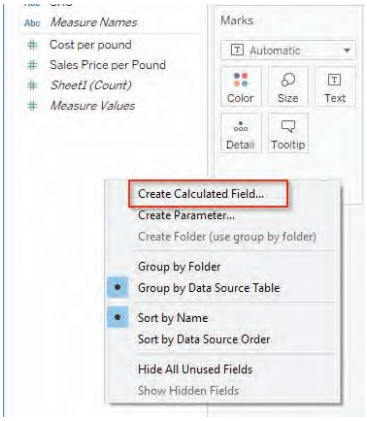
Tableau Software



Analyze the Data

Analysis Task 1: Compute the Gross Margin and Gross Margin Percentage for Each Product (SKU).

1. Right-click below the variable names and select Create Calculated Field . . . (Lab Exhibit 1.3T.2).



Lab Exhibit 1.3T.2

Tableau Software

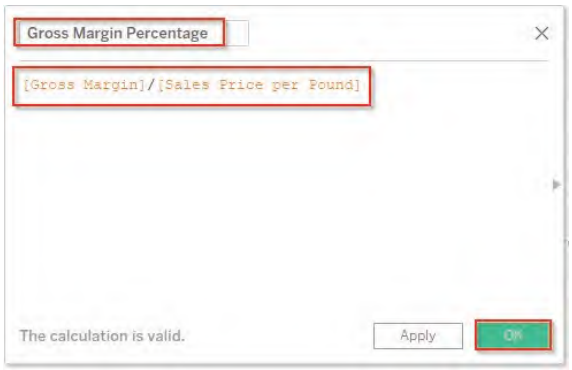
2. Name your new calculated field **Gross Margin.**
3. Gross Margin is calculated by subtracting the **Cost per Pound** field from the **Sales Price per Pound** field. Type in **Sales Price per Pound – Cost per Pound** (Lab Exhibit 1.3T.3). Tableau will automatically configure the words into references to the fields in the data set.



Lab Exhibit 1.3T.3

Tableau Software

4. Create a Second Calculated Field. Name your new calculated field **Gross Margin Percentage,** which is the percent of every dollar of sales that is gross margin. It is calculated by dividing Gross Margin by Sales Price per Pound. Type in "Gross Margin / Sales Price per Pound." Tableau will automatically configure the words into references to the fields in the data set (Lab Exhibit 1.3T.4). Click OK.



Lab Exhibit 1.3T.4

Tableau Software

Analysis Task 2: Create a Histogram.

1. To get started, we want to create a histogram to display the distribution of gross margin percentage across SKUs.
   Double-click **Gross Margin Percentage** to begin working with the measure.
   Notice that Tableau defaults to displaying the value as an aggregate measure (SUM) in a bar chart.
2. To change this default visualization to a histogram, select histogram from the **Show Me** menu, as shown in Lab Exhibit 1.3T.5.

Lab Exhibit 1.3T.5

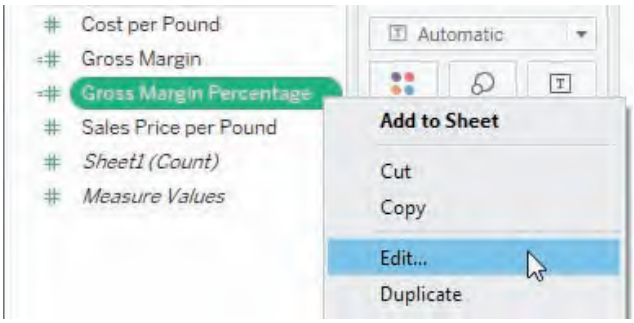
Tableau Software



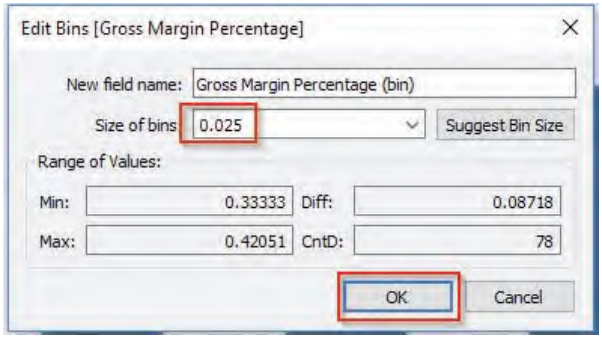
3. The histogram will default to very narrow bins.
4. To change the bin size, select the new **Gross Margin Percentage (bin)** dimension, and then select Edit from the menu (Lab Exhibit 1.3T.6).

Lab Exhibit 1.3T.6

Tableau Software



5. Change the Size of bins to 0.025, then click OK, as shown in Lab Exhibit 1.3T.7.

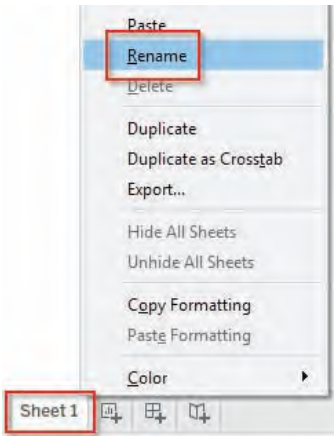


Lab Exhibit 1.3T.7

Tableau Software

Analysis Task 3: Access a List of SKUs in the Highest Bin.

1. We'll now work on the next requirement: Get a list of SKUs in the highest 0.025 bin (from 0.40 to 0.425). To do so, we can sort the gross margin percentages from lowest to highest and copy the 24 SKUs in that range.
   Right-click on **Sheet 1,** then select Rename to rename the sheet (Lab Exhibit 1.3T.8).



Lab Exhibit 1.3T.8

Tableau Software

2. Rename the sheet **Histogram.**
3. Open a new sheet, as shown in Lab Exhibit 1.3T.9.



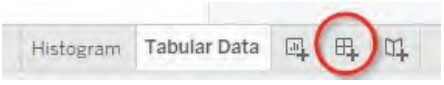
Lab Exhibit 1.3T.9

Tableau Software

4. Name this new sheet **Tabular data.**
5. Double-click **SKU** to get a unique list of all of the SKUs in the data set, and then double-click **Gross Margin Percentage.**
6. Double-click **Description** to make the tabular data more interesting and intuitive to interpret.

Analysis Task 4: Create a Dashboard Summarizing the Analysis Performed.

1. We can create a dashboard to view both of our charts in the same pane. Click the icon to create a new Dashboard (Lab Exhibit 1.3T.10).



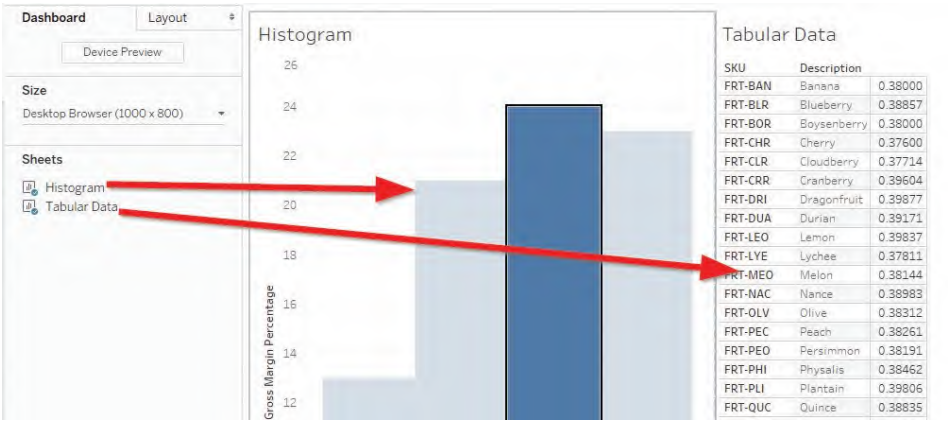
Lab Exhibit 1.3T.10

Tableau Software

2. Drag and drop each Sheet (Histogram and Tabular Data) to the Dashboard pane (Lab Exhibit 1.3T.11).

Lab Exhibit 1.3T.11

Tableau Software



3. Click the histogram sheet to make it active, and select the Use as Filter button. This will allow you to click any of the bars in the histogram to activate that bin as a filter for the entire dashboard (Lab Exhibit 1.3T.12).

Lab Exhibit 1.3T.12

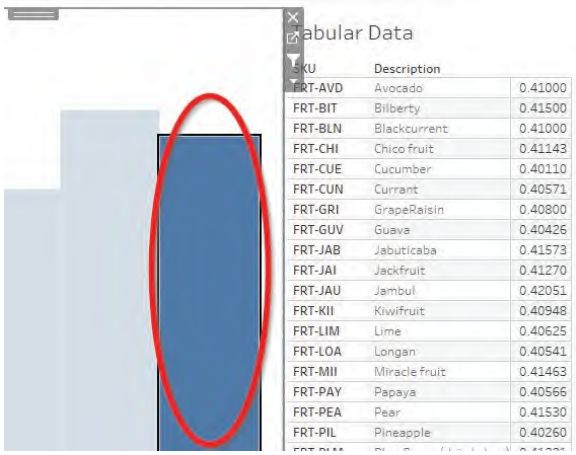
Tableau Software



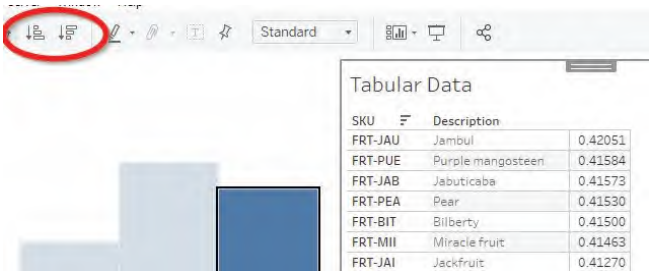
4. Click the bin with the highest percentages in it to filter the Tabular Data list (Lab Exhibit 1.3T.13).

Lab Exhibit 1.3T.13

Tableau Software



5. When in the tabular graphic, you can also sort your results by clicking the sort buttons in the menu (Lab Exhibit 1.3T.14).



Lab Exhibit 1.3T.14

Tableau Software

Report the Results

With the histogram and the dashboard, we are ready to show the results to management to help assess which products the company should promote because they are the most profitable.

Assessment

1. Take a screenshot of your histogram, paste it into a Word document named "Lab 1.3 Tableau Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of your dashboard, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 1.3 Tableau Submission.docx via Connect, if it is assigned.

Lab 1.3 Power BI

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Gross margin: Net sales minus the cost of goods sold. Gross margin is the amount of money a company retains after deducting the direct costs of producing a good or service.

SKU: Stock-keeping unit; a unique code for each product sold.

Assessing Product (SKU) Profitability

Keywords

Customer Profitability, Inventory Management, Customer Relationship Management

Decision-Making Context

Merchandising companies sell products to customers. They make money by selling the products for more than what they paid. In this lab, we compute the **gross margin** for each product to identify the company's most profitable products. This analysis allows merchandising companies to focus their efforts on selling the most profitable products, which in turn helps them make more profits.

Required

1. Create calculated fields for the gross margin and gross margin percentage.
2. Create a histogram to visualize the distribution of gross margin percentage across SKUs.
3. Adjust the automated bin size to 0.025.
4. Create a list of SKUs in the highest bin.

Specify the Question

Which products (SKUs) are the most profitable to sell?

Obtain the Data

Fancy Fruits is a (fictitious) mail-order business that sells exotic fruit from around the world to customers throughout the United States. The company keeps a price list for each product. We will work from this price list to find the items that are most profitable per dollar of sales. The data dictionary explains which data are included in the data set.

Data Dictionary

SKU:	Fancy Fruits' SKU (unique code for each fruit product sold)
Description:	Description of each fruit sold
Quantity (Pounds):	Weight of each product sold
Sales Price per Pound:	Total sales price per pound of product
Cost per Pound:	Total cost of product sold per pound of product

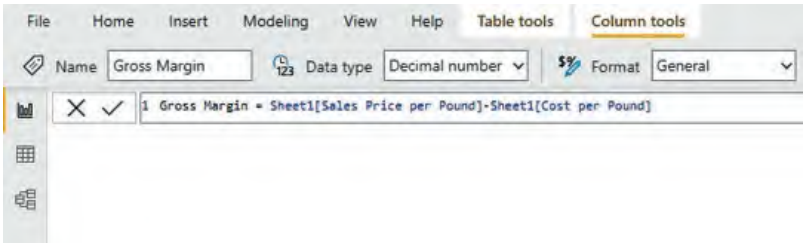
1. Open Power BI Desktop. From Power BI Desktop, select Home > Get Data > Excel Workbook, then browse to the location where you have saved the file Lab 1.3 Data.xlsx, and click Open.
2. In the Navigator window, check the box next to Sheet 1 and click Load to begin working with the data.

Analyze the Data

Analysis Task 1: Compute the Gross Margin and Gross Margin Percentage for Each Product (SKU).

1. In the Ribbon, select **Home > Modeling > New Column** to create a new calculated field.

2. Name your new calculated field **Gross Margin.** Gross Margin is calculated by subtracting the **Cost per Pound** field from the **Sales Price per Pound** field. Type in "GrossMargin = Sheet1[Sales Price per Pound] – Sheet 1[Cost per Pound]" (as shown in Lab Exhibit 1.3P.1).



Lab Exhibit 1.3P.1

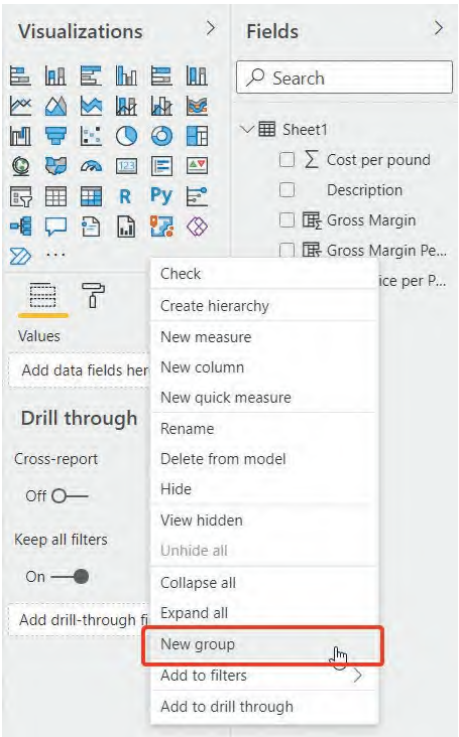
Microsoft Power BI

3. Create a Second Calculated Field. Name your new calculated field **Gross Margin Percentage,** which is the percent of every dollar of sales that is gross margin. It is calculated by dividing Gross Margin by Sales Price per Pound. Type in "Gross Margin Percentage = [Gross Margin]/Sheet1[Sales Price per Pound]".

Analysis Task 2: Create a Histogram.

1. To get started, we want to create a histogram to display the distribution of gross margin percentage across SKUs.
2. Right-click **Gross Margin Percentage** in your Fields Pane and select New group (as shown in Lab Exhibit 1.3P.2).

DATA VISUALIZATION



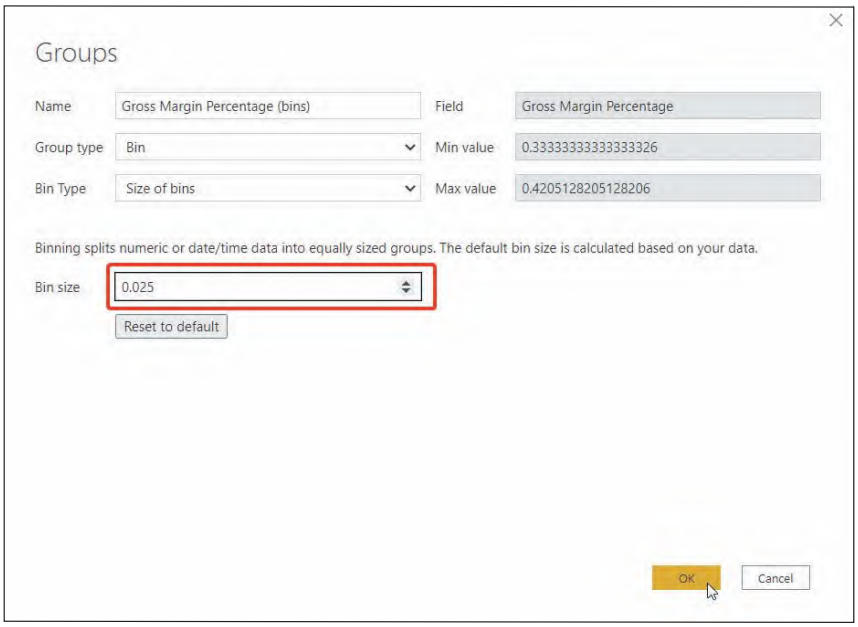
Lab Exhibit 1.3P.2

Microsoft Power BI

3. The histogram will default to very narrow bins. Change the Bin size to 0.025, and then click OK (as shown in Lab Exhibit 1.3P.3).

Lab Exhibit 1.3P.3

Microsoft Power BI



4. Next, create the histogram. Click the blank area on your page, and then click the Stacked Column Chart icon in the Visualizations pane. Resize the chart to fill the left side of the page.
5. Check the box next to Gross Margin Percentage (bins) to add it to the axis.
6. Check the box next to Gross Margin Percentage to add it to the Values. It will default to a Sum, so click the drop-down next to Gross Margin Percentage in the Values area and select Count to change the aggregate value.

Analysis Task 3: Access a List of SKUs in the Highest Bin.

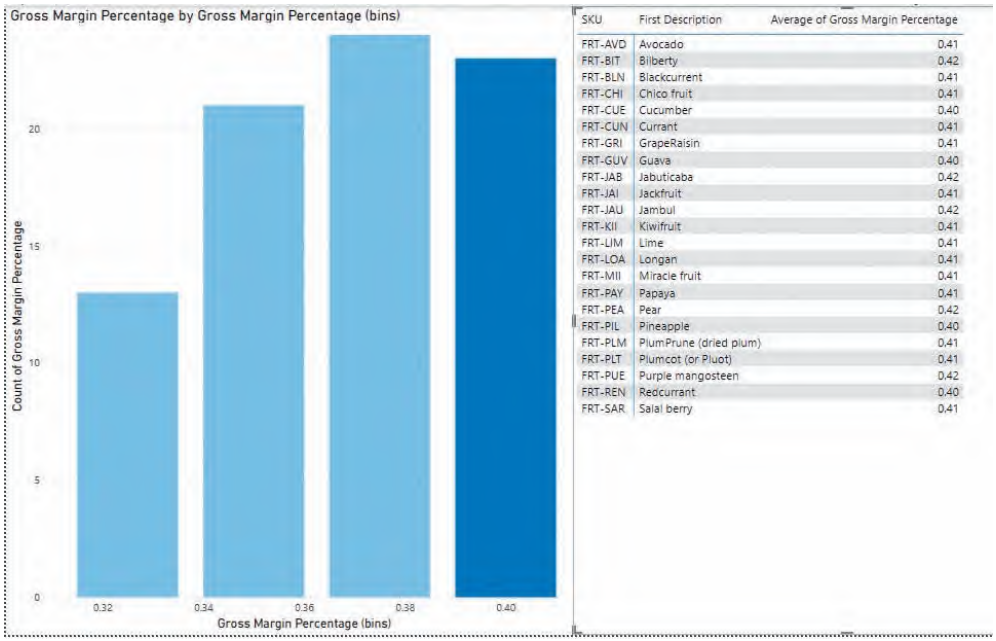
1. We'll now work on the next requirement: Get a list of SKUs in the highest 0.025 bin (from 0.40 to 0.425). To do this, we can sort the gross margin percentages from lowest to highest and copy the 24 SKUs in that range.
2. Click the blank area on your page, and then click the Matrix icon in the Visualizations pane. Resize the chart to fill the right side of the page.
3. Click SKU to get a unique list of all of the SKUs in the data set as rows.
4. Drag Description to the Values box to make the tabular data more interesting and intuitive to interpret.
5. Click Gross Margin Percentage to show the values.
6. Click the bin with the highest percentages in it to filter the Tabular Data list (as shown in Lab Exhibit 1.3P.4).
7. Remember that you can also sort your results by clicking the sort buttons in the menu.

Report the Results

We report the results to the sales staff and company management to identify the products (SKUs) that are most profitable. The company may want to emphasize the marketing of these products in the future.

Assessment

1. Take a screenshot of page 1 with the histogram and the table, paste it into a Word document named "Lab 1.3 Power BI Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of your dashboard, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 1.3 Power BI Submission.docx via Connect, if it is assigned.



SKU	First Description	Average of Gross Margin Percentage
FRT-AVD	Avocado	0.41
FRT-BIT	Bilberry	0.42
FRT-BLN	Blackcurrent	0.41
FRT-CHI	Chico fruit	0.41
FRT-CUE	Cucumber	0.40
FRT-CUN	Currant	0.41
FRT-GRI	GrapeRaisin	0.41
FRT-GUV	Guava	0.40
FRT-JAB	Jabuticaba	0.42
FRT-JAI	Jackfruit	0.41
FRT-JAU	Jambul	0.42
FRT-KII	Kiwifruit	0.41
FRT-LIM	Lime	0.41
FRT-LOA	Longan	0.41
FRT-MII	Miracle fruit	0.41
FRT-PAY	Papaya	0.41
FRT-PEA	Pear	0.42
FRT-PIL	Pineapple	0.40
FRT-PLM	PlumPrune (dried plum)	0.41
FRT-PLT	Plumcot (or Pluot)	0.41
FRT-PUE	Purple mangosteen	0.42
FRT-REN	Redcurrant	0.40
FRT-SAR	Salal berry	0.41

Lab Exhibit 1.3P.4

Microsoft Power BI

Alternate Lab 1.3

Assessing Product (SKU) Profitability

Apply the same steps in Lab 1.3 Excel (or Lab 1.3 Tableau or Lab 1.3 Power BI) to the Alt Lab 1.3 Data.xlsx data set using your tool of choice.

Open Excel file Alt Lab 1.3 Data.xlsx and follow similar directions to those in Lab 1.3 Excel above. ThrustMaster is a fictitious company that sells game controllers, joysticks, and steering wheels for PCs and for Xbox, Nintendo, and PlayStation consoles. A summary of the price list of its products by SKU is included in the Excel spreadsheet.

Required

1. Draw a histogram of the **gross margin** percentages using bins of 0.025 size. Use bins 0.425, 0.45, 0.475, and 0.5.
2. Get a list of SKUs in the highest 0.025 bin (from 0.425 to 0.50).



DATA VISUALIZATION

Gross margin: Net sales minus the cost of goods sold. Gross margin is the amount of money a company retains after deducting the direct costs of producing a good or service.

Assessment

1. Take a screenshot of your detailed listing of fruit products sold from the 0.40 to 0.425 gross margin percentage range, paste it into a Word document named "Alt Lab 1.3 Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of your histogram as a pivot chart, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 1.3 Submission.docx via Connect, if it is assigned.

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Chapter 2

Obtain the Data:
An Introduction to Business Data Sources

A Look Back

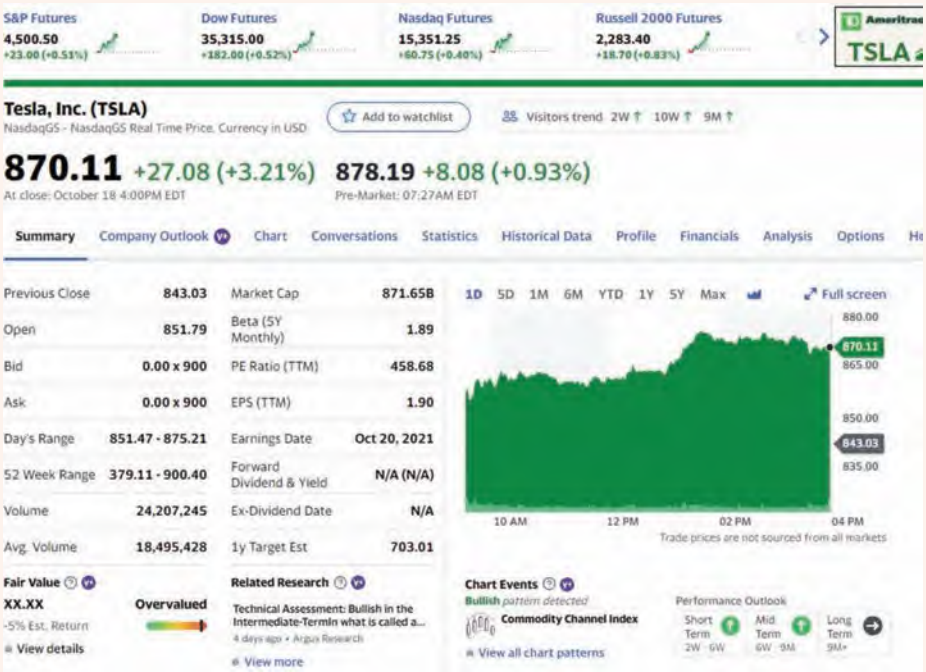
- Businesses, which exist to create value, collect and have access to an enormous amount of data.
- Business analysts use data to help managers answer questions about value creation and other business functions.
- The SOAR analytics model is a useful framework for performing business analytics and providing advice to managers.
- The first step of the SOAR model is SPECIFY THE QUESTION.

A Look at This Chapter

- To address business questions, we must know what data are available, how to access those data, and how to evaluate those data.
- Chapter 2 introduces the second step in the SOAR model: OBTAIN THE DATA. It summarizes the sources of business data, the various types of data, and key data management tools. It also explains how to access and prepare data for analysis.

A Look Ahead

- Chapter 3 introduces the statistical toolkit required for business analytics.
- Analysts use summary statistics and visualization methods to understand the distribution of data.
- Statistical tests, hypothesis tests, correlations, and regression analysis are important tools in the business analyst's toolkit.



Source: Yahoo Finance, Accessed 8/19/2020, https://finance.yahoo.com/quote/TSLA?p=TSLA&.tsrc=fin-srch.

Viewing Data in Reports versus Creating Reports Ourselves

Data come in many forms and from many sources. For example, the screenshot shown here provides data regarding **Tesla**'s stock price on a certain day and related data about the company's recent performance.

In any data analysis project, we must know the source of the data so that we can test those data for reliability. When we are presented with data in a financial report or some other aggregated form —for example, this screenshot from **Yahoo** Finance—it can be difficult to recognize the sources of the data and how the numbers were created. For example, we know that the source of this screenshot is Yahoo Finance, but where did Yahoo Finance get its numbers from?

In this chapter, we describe some of the many sources of data and how we might use them in business analysis.

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

- **LO 2.1** Identify common internal and external sources of data.
- **LO 2.2** Identify the two most common formats requested for data.
- **LO 2.3** Identify the different types of structured data.
- **LO 2.4** Understand how to prepare data for analysis.
- **LO 2.5** Identify common tools used to prepare data for analysis.
- **LO 2.6** Define data ethics and describe how to gather, protect, and use personally identifiable information in an ethical manner.

LABS FOR CHAPTER 2

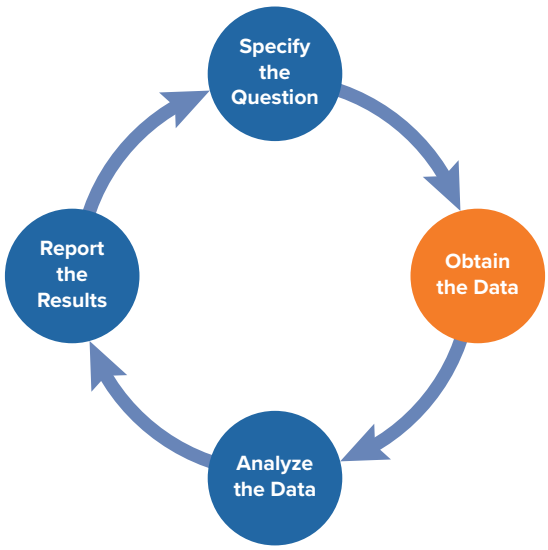
- **LAB 2.1** **Excel:** Identifying and Working with Different Data Types
- **LAB 2.2** **Tableau:** Preparing Different Data Types for Analysis
- **LAB 2.2** **Power BI:** Preparing Different Data Types for Analysis
- **LAB 2.3** **Tableau:** Conducting Preliminary Exploratory Analysis
- **LAB 2.3** **Power BI:** Conducting Preliminary Exploratory Analysis
- **LAB 2.4** **Excel:** Aggregating and Visualizing Different Data Types

In Chapter 1, we introduced the SOAR model (shown in Exhibit 2.1). The first step is to SPECIFY THE QUESTION. After we carefully specify the question, we need to look for data that can help answer it. In this chapter, we take a deep dive into the second step of the SOAR model: OBTAIN THE DATA. Business analysts take three steps to obtain the data they need:

1. Identify available, accessible data sources.
2. Obtain the data.
3. Prepare the data for analysis.

This chapter primarily explores the first step: identifying available data sources to use in business analysis. The current and increasing abundance of data along with technology provides business analysts with the opportunity to perform some of the same tasks as data scientists, and to increasingly rely on data in addition to experience, history, and intuition.

Exhibit 2.1
The SOAR Analytics Model



LO 2.1

Identify common internal and external sources of data.

2.1 INTERNAL AND EXTERNAL DATA SOURCES

Data come in many varieties and exist in many forms. Some data sets are messy, with missing or otherwise erroneous data; other data sets are complete, error-free, and ready for analysis. Some data are already housed within a company; these sources of data are called **internal data** sources. In contrast, **external data** sources are housed outside the company and will take some effort, and possibly money, to access.

Internal Data Sources

Companies frequently manage a single, large integrated database that is used to create, store, and maintain the data that are created during day-to-day business operations. This database typically includes data about sales, purchasing, accounts payable, manufacturing, customers, suppliers, inventory, and employees. Usually it is quite easy for authorized personnel to request the particular data needed for analysis and receive them in a desired format.

Internal databases can be relatively small and held on a desktop, or they can be quite large. When companies outgrow smaller databases, they may choose to implement an **enterprise resource planning (ERP)** system. Also known as an **enterprise system**, an ERP system is a type of business management software that integrates applications from throughout the business (such as manufacturing, accounting, finance, and human resources) into one system. ERP systems are categorized as **centralized databases** because they store all of a company's internal, structured data in one location.

Several different vendors provide ERPs; the most popular are **SAP**, **Oracle**, and **Workday**. ERPs are useful because their integration of a company's various systems provides "one version of the truth." In contrast, if a company manages its operations across a variety of isolated databases, it is possible that product codes, quantities on hand, and the formatting of financial values (for example, with different types of rounding) could be inconsistent across systems and lead to inaccurate reporting and poor decision-making.

Not everyone in the company is permitted to access, edit, and/or create data across the entire integrated system. For example, certain employees may be permitted to work only within certain screens, as when a customer sales representative is limited to working within the Sales Order functions to view past sales and to help answer customers' questions. A customer sales representative would not be able to view any data related to purchasing inventory, accounts payable, or employee compensation.

Retrieving Data from Centralized Databases

As we've seen, internal data are often stored in a centralized database. Centralized **relational databases** compile the data into separate tables, with each table containing a unique list of the items stored. Relational databases are means of storing data to ensure that the data are complete and not redundant. They help enforce business rules, such as which employees are allowed to perform different job duties, the maximum limit for a sales order before it needs management approval, or the number of payments accepted for a single sales transaction. They also aid in communication and integration of business processes across an organization.

A relational database is composed of tables, fields, and records.

- **Tables** are data organized into sets of columns (fields) and rows (records).
- **Fields** (also called *attributes* or *variables*) are the columns that contain descriptive characteristics about the observations in the table.
- **Records** are the rows, with each observation corresponding to a unique instance of what is being described in the table.

Any field that functions as a unique identifier in a table has a special name, called a **primary key**. In relational databases, every table must have a primary key because every table represents a unique listing of the data stored in it.

For example, in one relational database, a Customers table contains details about each customer, such as their name, address, and phone number, along with a unique Customer ID. A separate Sales Transaction table contains a listing of every transaction between the company and its customers along with a unique identifier, Transaction ID. To identify which customer was involved in each transaction, there must be a common field between the two tables, in this instance the Customer ID. This field *relates* the Transaction table to the Customers table so that given the transaction, customer information can be retrieved, as shown in Exhibit 2.2.

The color-coding and lines drawn between the two tables show how the tables interact. In this instance, CustomerID is the field that is common between the two tables. So, in the Customers table, CustomerID is the *primary key* because it functions to uniquely identify every customer in the table. The CustomerID field also exists in the Sales Transaction table so that the two tables can be linked.

In the Transaction table, Customer_ID is not uniquely identifying anything. Instead, it is creating a look-up relationship. So, in this case, the field Customer_ID is a foreign key. **Foreign keys** exist to create relationships between two tables so that database users can look up details of the observation based on the primary key/foreign key relationship.

Looking at Exhibit 2.2, we can determine that Paula participated in two transactions on January 1 (one was a return, one was a sale). Terry participated in three different transactions on January 1, and Edna participated in a sale on January 2.

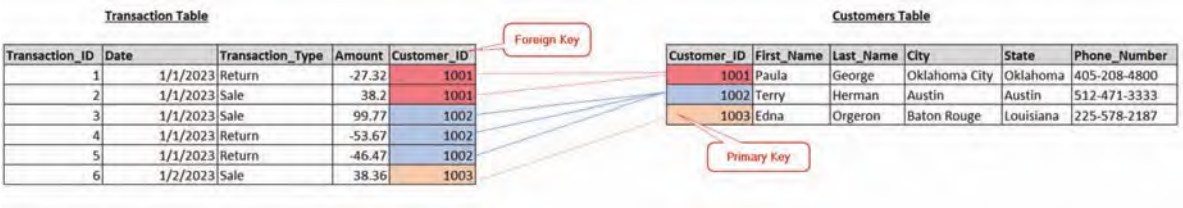


Exhibit 2.2 Relational Database: Primary Key and Foreign Key

Unlike primary keys, foreign keys are not required in every table. However, when a foreign key exists, it must contain matching data in the related table. Its purpose is to allow users to look up related data, so if there are no related data, a foreign key isn't useful. In other words, if we assume that the Customers table in Exhibit 2.2 is a complete list of customers, then we would not be able to include any Customer_ID in the Transaction table that was not 1001, 1002, or 1003. If we needed to record a new customer on a new transaction, we would have to record the data in the Customers table first. The new Customer can then be associated with the next sequential number in the primary key field, Customer_ID 1004, and we could then add Customer_ID 1004 to a transaction in the Transaction table.

PROGRESS CHECK

1. How do internal and external data sources differ? Which is easier to access? With which type of data are business analysts likely to be familiar?
2. Why is it important to keep some data strictly internal to a company (and not provided to people outside of the company)?
3. Why is it beneficial to store all internal company data in one integrated system?
4. Why are some users of an ERP system permitted only to view the data, and not to edit it directly?
5. What is the difference between a primary key and a foreign key?

External Data Sources

All companies must consider the world around them before rolling out new products, expanding their operations, or downsizing. In addition to using their own operational data for business analysis, companies can access external data to provide additional context for decision-making. Here are some examples of useful external data sources.

- *Social media data.* Many companies today reach customers via social media, and their marketing departments want to see data on audience engagement (likes/shares), click-throughs, and behavior patterns (such as days and times of highest and lowest engagement). Marketers can put the social media information into a dashboard (like the one in Exhibit 2.3) to help them determine if they are posting information when their target audience is most active across different social media platforms. **HootSuite** and **Google Analytics** are two popular applications for performing social media analytics. They show social media activity across the platforms on one page. In addition, analysts can use a programming language such as Python or R to retrieve data from the different social media sites—for example, to retrieve all of the comments on their company's public page (or a competitor's page) or to write content to the social media sites.

  Social-media management software, including HootSuite and Google Analytics, is built on APIs (application programming interfaces), which are tools used to help exchange data

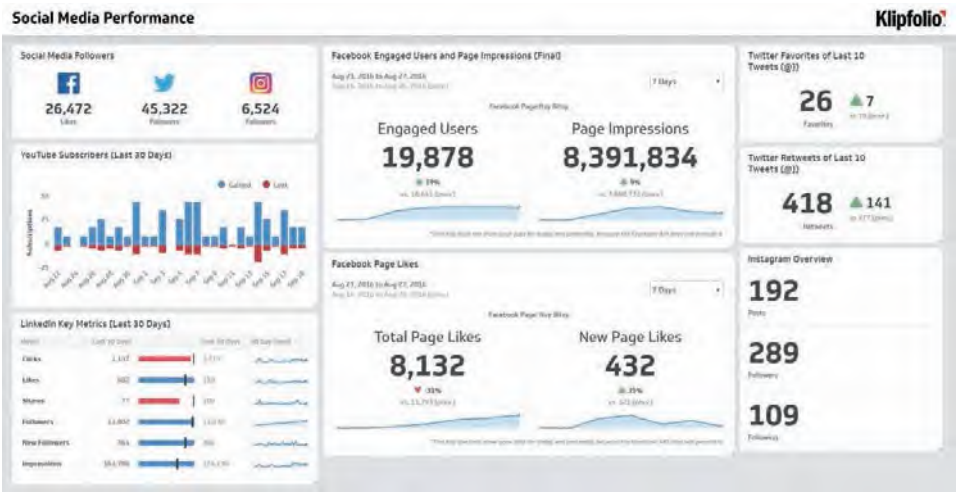


Exhibit 2.3
Example of Social Media Performance Blackboard

(Source: Klipfolio.)

between different software. If a company uses multiple social media platforms such as **Twitter**, **Facebook**, and **Instagram**, to reach clients or customers, the software can use the API connections to show social media activity across the platforms on one page.

- *U.S. Census Bureau data.* The U.S. Census Bureau provides a tremendous amount of information about the U.S. population. As just one example, businesses can use these data to inform decisions regarding potential markets and where they should open a new location. Exhibit 2.4 shows just a small sample of the data that are available on the U.S. Census Bureau's website (data.census.gov/mdat). We have expanded one of the variables so that you can see an example description and the different values for one particular variable, "Gross rent as a percentage of household income past 12 months."



Exhibit 2.4 Data Available from the U.S. Census Bureau

(Source: data.census.gov/mdat.)

- *Small Business Administration (SBA) data.* The data and resources compiled and provided by the U.S. Small Business Administration (SBA) can provide valuable insights regarding the current state of the economy, population trends, and small business conditions. The following links are particularly useful:
  - https://www.sba.gov/business-guide/plan-your-business/market-research-competitive-analysis provides guidance on market research and competitive analysis.
  - https://www.sba.gov/about-sba/sba-performance/open-government/digital-sba/open-data/open-data-sources provides links to the SBA data resources.

  Exhibit 2.5 shows a sample of the data that the SBA has collected.

Exhibit 2.5
Small Business Administration Data

(Source: https://www.sba.gov/business-guide/plan-your-business/market-research-competitive-analysis.)

Free small business data and trends

There are many reliable sources that provide customer and market information at no cost. Free statistics are readily available to help prospective small business owners.

Consider these types of business statistics in your market research and competitive analysis:

Focus	Goal	Reference
General business statistics	Find statistics on industries, business conditions	NAICS, USA.gov Statistics, Statistical Abstract of the United States, U.S. Census Bureau
Consumer statistics	Gain info on potential customers, consumer markets	Consumer Credit Data, Consumer Product Safety
Demographics	Segment the population for targeting customers	American FactFinder, Bureau of Labor Statistics
Economic indicators	Know unemployment rates, loans granted and more	Consumer Price Index, Bureau of Economic Analysis
Employment statistics	Dig deeper into employment trends for your market	Employment and Unemployment Statistics
Income statistics	Pay your employees fair rates based on earnings data	Earnings by Occupation and Education, Income Statistics
Money and interest rates	Keep money by mastering exchange and interest rates	Daily Interest Rates, Money Statistics via Federal Reserve

- *Publicly available financial data, such as stock performance data, summarized financial data, and financial statements.* Financial data help business analysts assess not only how their company is performing compared to its own past performance, but also how it is performing compared to its peers and competitors. Publicly available data sources can also help companies understand Wall Street's valuation of their firm and help them evaluate potential investments. The **Securities and Exchanges Commission (SEC)**, which is the U.S. government agency charged with protecting investors by requiring full and truthful financial disclosure by companies, keeps a repository of financial statement information on its EDGAR website (https://www.sec.gov/edgar.shtml). **Compustat** also maintains a database of the financial statement items. **Yahoo! Finance** (https://finance.yahoo.com/) maintains multiple years and quarters of financial statement data and stock price data.

Big Data Sources

In the previous sections, you learned about data sources that are considered more traditional and structured. In recent years, businesses and individual consumers have captured a massive amount of unstructured data through mapping apps on cell phones, social media interactions, and blog comments.

The term **Big Data** refers to data sets that are too large and complex for businesses' centralized systems to capture, store, manage, and analyze. Big Data is characterized by the four Vs: volume, variety, velocity, and veracity (Exhibit 2.6).

Volume refers to the sheer amount of data, regardless of its source. Indeed, Eric Schmidt, CEO of Google, says, "Every two days we now create as much information as we did from

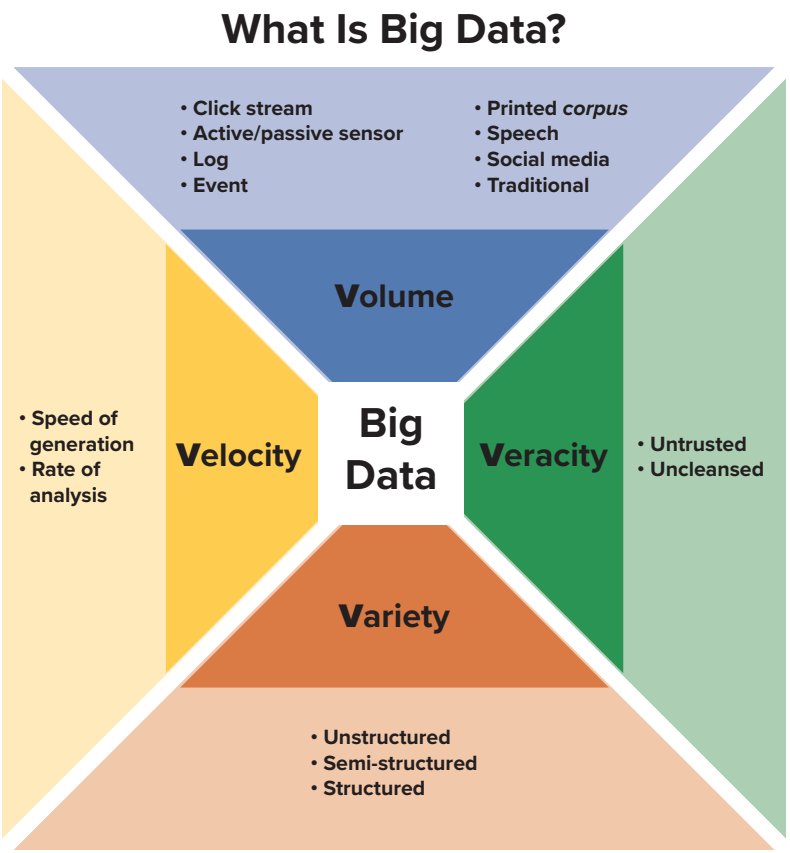


Exhibit 2.6
The Four Vs of Big Data: Volume, Variety, Velocity, and Veracity

(Source: EY, "Big Data: Changing the Way Businesses Compete and Operate," *Insights on Governance, Risk, and Compliance,* April 2014, p. 2. Available online at https://motamem.org/wp-content/uploads/2019/01/Big-data-applications-and-insights.pdf.)

the dawn of civilization up until 2003."[1] The data might come from corporate systems, social media, the government, speeches, company press releases, or internet searches. Sorting through these mounds of data to find the data that are relevant to a decision is often one of the biggest challenges facing data analysts.

Variety refers to the different forms of data. **Structured data** are highly organized data that fit neatly in a table or a database. **Unstructured data** are data without internal organization or structure. Blogs, tweets, and pictures posted to Instagram are examples of unstructured data. Some people estimate that 80 percent of enterprise data (including emails, PDFs, and other documents) are unstructured.[2] Between those two extremes is **semi-structured data**, which has elements of both structured and unstructured data. Although semi-structured data do not have any internal structure (such as labeled columns), they may come with tags or markers explaining what the data represent. HTML, which is the programming language that supports the creation of webpages, is an example of semi-structured data. So are emails that have tags such as recipient, sender, date, subject, and so on. The **Internet of Things (IoT)** is an example of the increasing variety and volume of data. It is composed of physical objects (such as cars, watches, and refrigerators)

[1]M. G. Siegler, "Eric Schmidt: Every 2 Days We Create as Much Information as We Did up to 2003." *TechCrunch,* https://techcrunch.com/2010/08/04/schmidt-data/. Accessed July 1, 2019.
[2]J. Rizkallah, "The Big (Unstructured) Data Problem," *Forbes,* 2017, https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/#532a74a4493a. Accessed November 16, 2018.

that have sensors connecting them to the internet and allowing for the exchange of data. IoT data are a mix of structured, semi-structured, and unstructured data.

Velocity refers to the speed that the data are being generated or the rate that data are being analyzed. For example, stock prices, which can change from one second to another, are generated and analyzed much more frequently than a company's financial statements, which are generated monthly, quarterly, or annually.

Veracity refers to the underlying truthfulness, accuracy (cleanliness), and trustworthiness of data. Data that are untrustworthy, irrelevant, inconsistent, or biased can and do lead to poor decisions.[3]

Traditionally, Big Data was too large for on-site storage and too varied to function in a traditional centralized database. However, there now are several options for storing and working with Big Data, including data warehouses, data lakes, and distributed computing.

For companies that do not have one centralized database, a **data warehouse** can integrate different databases across the company. For instance, if different teams or different departments maintain their own separate databases, a data warehouse can make the company's data storage more efficient by integrating the disparate data sources in one large storage location. In addition, companies that want to integrate external data sources with their internal data can compile all those data in a data warehouse. There is just one caveat: Data warehouses store and process only structured data, not unstructured data.

Like a data warehouse, a **data lake** integrates data from different sources, but it can maintain much larger amounts of data, including both structured and unstructured data. It may seem that a data lake is always preferable to a data warehouse because of the added flexibility that comes with maintaining more data, including unstructured data. However, data lakes do have an important limitation: The more data that are stored, the greater is the risk of delayed reporting and information overload. A data lake can make it more difficult for business analysts to focus on precisely the data they need to perform their analysis.

Data warehouses and data lakes store and integrate data, but remember that analysts must work with data, not just collect data. A variety of programming languages can be used to retrieve the data, such as SQL, R, and Python, and a processing application called an OLAP cube can also be built on top of a data warehouse. **OLAP (online analytical processing)** is a computing method that enables users to easily and selectively extract and query data for analysis from different points of view. **OLAP cubes** are a data structure of compressed, multidimensional data for faster analytics processing (that is, all possible combinations are pre-calculated because speed is the number-one goal). OLAP cubes are typically created for different functions in the company, allowing their use in visualizations and dashboards.

However, if a company is using a data lake to store unstructured data in addition to structured data and an analyst wishes to perform analysis on the unstructured data, then OLAP may not be useful. In this instance, an upgrade to distributed computing is recommended. In **distributed computing**, data are stored across multiple databases, which allows greater storage of data and increased speed of analysis.

An example of a framework that supports distributed data lakes is **Hadoop**, an open-source framework for storing and processing Big Data. Hadoop was a trailblazing framework for Big Data, and many organizations continue to use it to work with Big Data. For a company to acquire the computation power necessary to run Hadoop (or any other big data framework) with massive data sets, it must shift to cloud storage. Cloud storage vendors such as **Amazon Web Services (AWS)**, **Cloudera**, and **Microsoft Azure** provide cloud solutions to support the storage of massive amounts of data. All are compatible with Hadoop.

[3]Richardson, V. J., & Watson, M. W. (2021). Act or Be Acted Upon: Revolutionizing Accounting Curriculums with Data Analytics. *Accounting Horizons, 35*(2), 129-144.

PROGRESS CHECK

6. Which types of data that are internal to the company would be useful in preparing a sales forecast for the next quarter? Which types of data that are external to the company would be meaningful for the same purpose?
7. How can a business analyst use data from the U.S. Census Bureau to understand the demographics of a company's customer base?
8. How can companies improve their traditional business analytics by including Big Data sources such as social media or data from the Internet of Things?

2.2 OBTAINING THE DATA

LO 2.2

Identify the two most common formats requested for data.

Whether data are stored internally or purchased from an external source, analysts need to obtain or request data in a specific format. The two most commonly requested data formats are text and tabular.

Text Data

Business analysts who plan to analyze mostly unstructured data that are text-based, such as customer reviews, may prefer to receive data as a text file. Common text editors are **Microsoft** Word, **Adobe** PDF, and Notepad. Data requested in text format will likely be provided as a generic text file (file extension .txt) so that the file is platform-independent. Other examples of data that are best stored in a text format include transcripts of phone calls and social media data (collections of tweets, hashtags, **Facebook** status updates).

Tabular Data

When you think of data, you likely think of structured tabular data. **Tabular data** are structured into rows and columns, with each column dedicated to one attribute of what is being described overall in the table. For example, a data set describing a company's employees would include one column for employee last name, another column for employee first name, another column for the employee's role in the company, and so on. Each row corresponds to one instance of the data. Continuing with the employee data example, each row would represent one individual employee identified using a unique employer ID number. Tabular data include the same number of cells for each row (equal to the number of columns), although some cells may be empty if certain data do not exist or have not been collected.

With tabular data, each cell should contain only a single value, and never multiple values. For instance, if an employee has multiple email addresses, they should not all be entered in a single cell. Rather, only one email address should be stored, or the data should be organized to allow multiple fields for email (for example, Preferred Email Address and Secondary Email Address). Exhibit 2.7 shows an example of employee data stored in a table.

EmployeeID	LastName	FirstName	Email	Title	BirthDate	HireDate	Address
1	Davolio	Nancy	ndavolio@nort	Sales Represer	12/8/1948	5/1/1992	507 - 20th Ave.
2	Fuller	Andrew	afuller@northv	Vice President	2/19/1952	8/14/1992	908 W. Capital
3	Leverling	Janet	jleverling@nor	Sales Represer	8/30/1963	4/1/1992	722 Moss Bay B
4	Peacock	Margaret	mpeacock@no	Sales Represer	9/19/1937	5/3/1993	4110 Old Redm
5	Buchanan	Steven	sbuchanan@nc	Sales Manager	3/4/1955	10/17/1993	14 Garrett Hill
6	Suyama	Michael	msuyama@nor	Sales Represer	7/2/1963	10/17/1993	Coventry Hous
7	King	Robert	rking@northwi	Sales Represer	5/29/1960	1/2/1994	Edgeham Hollc
8	Callahan	Laura	lcallahan@nort	Inside Sales Co	1/9/1958	3/5/1994	4726 - 11th Ave

Exhibit 2.7
Example of Tabular Data

When a project requires mostly structured data, requesting the data in a tabular format is likely the best option. Common applications for working with tabular data are Excel and **Google** Sheets. When data are requested in a tabular format, they are usually delivered in a comma-separated file (with the file extension .csv). This format keeps the data platform-independent and makes it easier to work with the data across a variety of tools and programming languages, such as Python or R. Additionally, .csv files do not have the same row and column limits that Excel has, which allows .csv files to hold Big Data that exceed Excel's size limit. Technically, .csv files store data as text, but they convert rapidly to a tabular format when they are imported into spreadsheet applications or used with programming languages.

Whether the data you intend to analyze are mostly numeric or mostly text, business analysis will be easier if the data are unformatted. Often, the data presented on websites and in reports have already been organized and formatted to make it easier for audiences to read and understand. But because you want to create your own insights, you should begin your analysis with data that have not already been aggregated or processed. The term **aggregated data** refers to individual data points that have been combined into subtotals, such as counts, sums, or averages. The benefit of working with **raw data** (data that have not been processed, cleaned, or aggregated) is the ability to select the most relevant data. If the data have already been cleaned, you lose the ability to see the data set in its entirety, and you will not see any data that were removed. When you clean the data yourself, you give yourself the opportunity to identify outliers and anomalies and to frame the data in ways that will best support decision-making.

All the data files used in this textbook are provided in either .xlsx (Microsoft Excel) or .csv format.

PROGRESS CHECK

9. Why are data provided in generic file formats such as .csv and .txt preferable to data provided in .xlsx or .docx formats?
10. Why are raw data preferable to aggregated data?

LO 2.3

Identify the different types of structured data.

2.3 STRUCTURED DATA TYPES: CATEGORICAL VERSUS NUMERICAL

To conduct business analytics, you must understand the data types you are working with. The data type helps you understand which types of analysis you can perform. In addition, specific data-visualization techniques are used for the different data types. You will continue to learn about data visualization in the later chapters of this textbook.

The two broadest categories of structured data are categorical and numerical. **Categorical data** are categorized into groups that are represented either by words or by non-meaningful numerical data, such as classifying a group of people by age (for example, adult, non-adult), labeling transaction types (for example, sales versus returns), or specifying inventory costing method (for example, FIFO, LIFO, average cost). In contrast, **numerical data** are meaningful numbers that represent quantities, such as transaction amount, net income, age, or score on an exam.

It is important to note that sometimes data can be presented as a string of numbers that do not have a true numerical value. Examples include your Social Security number, your driver's license number, and your Student ID number. Anytime we use a string of numbers to uniquely identify an item in a list, that attribute is considered categorical data because

the numerical aspect of the data is not meaningful. Instead, its sole purpose is to uniquely identify an item.

Categorical Data: Nominal and Ordinal Data

Let's now look more closely at the two types of categorical data: nominal and ordinal.

Nominal Data

The term **nominal data** refers to categorical data that cannot be ranked, such as country of origin and transaction type. The two primary methods to summarize categorical nominal data are

- Counting and grouping
- Proportion

Exhibit 2.8 is a snippet of a transaction table that includes both online and in-person sales. (This data set is a modified version of the data set in Lab 1.2.) Because each column represents a variable and each row represents a single transaction, we classify these data as structured. There is a mix of categorical data and numerical data in this data set. The first five columns—Date, Invoice #, Online or In-Person? Customer #, and SKU—are all categorical data. The remaining two columns, Sales Price and Cost, are numerical. Some of the categorical variables are more obviously categorical, such as Online or In-Person and SKU. These stand out as categorical because they are text strings. Date and Invoice # are also categorical even though they appear as strings of numbers, because these number strings do not represent mathematically meaningful numbers. We would have no reason to take an average of Invoice #, for instance. As mentioned earlier, categorical data can be represented by numbers as long as those numbers serve to identify an aspect of the data and do not have true numerical significance.

Date	Invoice #	Online or In-Person?	Customer #	SKU	Sales Price	Cost
10/1/2023	5001	Online	2506	PED-TPR	497.38	284.93
10/2/2023	5002	In-Person	2522	WHL-T3S	159.99	92.63
10/2/2023	5003	Online	2596	WHL-TGT	614.91	354.66
10/3/2023	5003	In-Person	2596	FLI-TMX	159.84	87.3
10/3/2023	5004	In-Person	2580	FLI-USB	21.99	12.06
10/3/2023	5005	In-Person	2555	WHL-T30	319.99	170.61
10/3/2023	5006	In-Person	2576	WHL-TMX	249.99	147.62
10/3/2023	5007	In-Person	2566	FLI-16H	97.3	52.53
10/3/2023	5007	In-Person	2566	FLI-HOX	39.99	20.53
10/3/2023	5008	In-Person	2595	FLI-16M	117.83	69.36
10/3/2023	5009	In-Person	2551	WHL-TGT	614.91	354.66
10/3/2023	5009	In-Person	2551	PED-FCS	39.19	23.36
10/3/2023	5010	In-Person	2566	FLI-HO1	59.99	30.17

Exhibit 2.8
Transaction Table Showing Online and In-Person Sales

Look more closely at the Online or In-Person column, which presents categorical, nominal data. We can count and group the transactions by the two categories represented by transaction type, as shown in Exhibit 2.9. Note that this table contains 13 rows or transactions. By highlighting the cells associated with online transactions, we can count how many transactions are online (2), which means the remaining 11 transactions are in-person transactions.

We can take this analysis of categorical data even further by calculating the proportion. We calculate **proportion** by taking the number of observations in a category and dividing that number by the grand total of observations in the sample.

Here, we calculate the relative proportion of online transactions (2) to the total number of transactions (13) by dividing 2 by 13, which shows that 0.15, or 15 percent, of the transactions in this table were online. By inference, we can determine that the 85 percent of remaining transactions (= 100% – 15%) took place in-person.

Exhibit 2.9
Analysis of Categorical Data (Online vs. In-Person Transactions)

Date	Invoice #	Online or In-Person?	Customer #	SKU	Sales Price	Cost
10/1/2023	5001	Online	2506	PED-TPR	497.38	284.93
10/2/2023	5002	In-Person	2522	WHL-T3S	159.99	92.63
10/2/2023	5003	Online	2596	WHL-TGT	614.91	354.66
10/3/2023	5003	In-Person	2596	FLI-TMX	159.84	87.3
10/3/2023	5004	In-Person	2580	FLI-USB	21.99	12.06
10/3/2023	5005	In-Person	2555	WHL-T30	319.99	170.61
10/3/2023	5006	In-Person	2576	WHL-TMX	249.99	147.62
10/3/2023	5007	In-Person	2566	FLI-16H	97.3	52.53
10/3/2023	5007	In-Person	2566	FLI-HOX	39.99	20.53
10/3/2023	5008	In-Person	2595	FLI-16M	117.83	69.36
10/3/2023	5009	In-Person	2551	WHL-TGT	614.91	354.66
10/3/2023	5009	In-Person	2551	PED-FCS	39.19	23.36
10/3/2023	5010	In-Person	2566	FLI-HO1	59.99	30.17

Ordinal Data

Like nominal data, **ordinal data** allow you to count, group, and calculate proportions based on the variable, but these data have one additional characteristic: There is a natural order to ordinal data that allows them to be ranked and sorted. There are three primary methods of summarizing ordinal data:

- Counting and grouping
- Proportion
- Ranking (or sorting)

Examples of ordinal data are letter grades (A, B, C, D, and F) and Olympic medals (gold, silver, bronze). Each of these variables can be counted (the number of As in a class, the number of gold medals won by a given country in the Olympic swimming events), and proportions can be calculated just as with nominal data. In addition, there is a natural order—we recognize that an A is associated with a higher grade than a B or a C. We also recognize that a gold medal is associated with a stronger performance than a silver or bronze. This characteristic of ordinal data allows you to sort your summary data in the natural order that the data present.

Ordinal data also allow sorting by rank. **Ranking** refers to a position on a scale. Ranking records based on the natural order of occurrence may be more informative than just numerical or alphabetical sorting because it provides added value to the analysis. For example, look again at the transaction table in Exhibit 2.9. We can interpret the Date variable as ordinal data, because the way the data are presented is not numerical, but there is a natural order associated with time. We can count the number of transactions associated with each date, but there is also added value in knowing not only the number of transactions by date, but also if the number of transactions increases or decreases over time. In short, the order of the dates is meaningful.

In Exhibit 2.10a, the different dates are grouped (highlighted), which makes summary analysis easier. We can quickly see that there was one transaction on October 1st, two on October 2nd, and so on. Knowing that there are 13 transactions total, we can calculate the proportion of transactions for each date. We can then create a summary table that shows the number of transactions on each date, as well as their respective proportions, as shown in Exhibit 2.10b. Because this table summarizes ordinal data, it is sorted by the Date variable.

Exhibit 2.10
Count of Transactions by Date

Date	Invoice #	Online or In-Person?	Customer #	SKU	Sales Price	Cost
10/1/2023	5001	Online	2506	PED-TPR	497.38	284.93
10/2/2023	5002	In-Person	2522	WHL-T3S	159.99	92.63
10/2/2023	5003	Online	2596	WHL-TGT	614.91	354.66
10/3/2023	5003	In-Person	2596	FLI-TMX	159.84	87.3
10/3/2023	5004	In-Person	2580	FLI-USB	21.99	12.06
10/3/2023	5005	In-Person	2555	WHL-T30	319.99	170.61
10/3/2023	5006	In-Person	2576	WHL-TMX	249.99	147.62
10/3/2023	5007	In-Person	2566	FLI-16H	97.3	52.53
10/3/2023	5007	In-Person	2566	FLI-HOX	39.99	20.53
10/3/2023	5008	In-Person	2595	FLI-16M	117.83	69.36
10/3/2023	5009	In-Person	2551	WHL-TGT	614.91	354.66
10/3/2023	5009	In-Person	2551	PED-FCS	39.19	23.36
10/3/2023	5010	In-Person	2566	FLI-HO1	59.99	30.17

(a)

Date	Count of Transactions by Date	Proportion of Transactions across each Date
10/1/2023	1	8%
10/2/2023	2	15%
10/3/2023	10	77%
Grand Total	**13**	**100.00%**

(b)

PROGRESS CHECK

11. What is the difference between nominal and ordinal data?
12. What is an example of nominal data? Of ordinal data?

Numerical Data: Interval and Ratio Data

Recall that numerical data are meaningful numbers, which means you can perform meaningful numerical analysis on them, such as summing, multiplying, or dividing. In numerical data, there is an equal interval (distance) between each observation, and each observation is measured along a scale. The four primary methods of summarizing numerical data are:

- Counting and grouping
- Proportion
- Summing
- Averaging

Numerical data can be divided into two subcategories: interval data and ratio data. In **interval data**, there is an equal and definitive interval or distance between each data point in the scale, but there is not a meaningful zero. The lack of a meaningful zero means that the value of zero does exist in interval data, but it does not represent the absence of anything—it is simply another number. A good example of interval data is temperature. Even though the difference of 1 degree on the Fahrenheit temperature scale is the same from 30 to 31 as it is from 77 to 78, which makes this data type numerical, a temperature of 0 degrees does not mean "the absence of temperature." Rather it is simply 1 degree below 1 degree and 1 degree above –1 degree. Another example of interval data is SAT scores. A student cannot even earn a zero on the SAT—the range of possible scores starts at 400 and ends at 1600, so 0 is certainly meaningless.

The majority of numerical data relating to business decisions is **ratio data**, which are the most sophisticated form of data. Like interval data, ratio data can be counted, ranked,

and used to perform numerical operations because the distance between each data point is equal and definitive. However, in ratio data the zero is meaningful, which allows for the calculation of ratios. A **ratio** is the numerical relation between two variables. Ratios are used to measure a company's financial position and identify areas for improvement. Because ratios are calculated with division, ratios cannot be calculated using interval data. Data measuring money—such as transaction amounts, expenses, revenues, assets, salary, and taxes—are all examples of ratio data.

Returning to our example in Exhibit 2.10, we can calculate meaningful numerical values with the attributes Sales Price and Cost. For example, we can calculate a grand total sum of Sales Price by adding each transaction amount together ($497.38 + $159.99 + $614.91, and so on) for a total of $2,993.30. We can work with these totals in a variety of ways, including slicing the totals by categorical attribute "Online or In-Person" to determine the total sales amount for each different transaction type.

MINI CASE: Tide PODS and Data Types

How are data types identified and included in a product database like the one **Amazon** maintains? Exhibit 2.11 is a screenshot from Amazon showing a listing for Tide PODS. It identifies the type of data for each field and suggests how each data type could be stored in a structured table.

As you read the different callout boxes, you will recognize terms that you learned earlier in the chapter, including numerical data and categorical data. You will also notice a new term: flag. Two data types in Exhibit 2.11 are denoted as a **flag**, which is a term used when there are only two options for a given field. Here, Amazon denotes two database items as flags: whether an item is prime or not, and whether the item is currently in stock or out of stock.

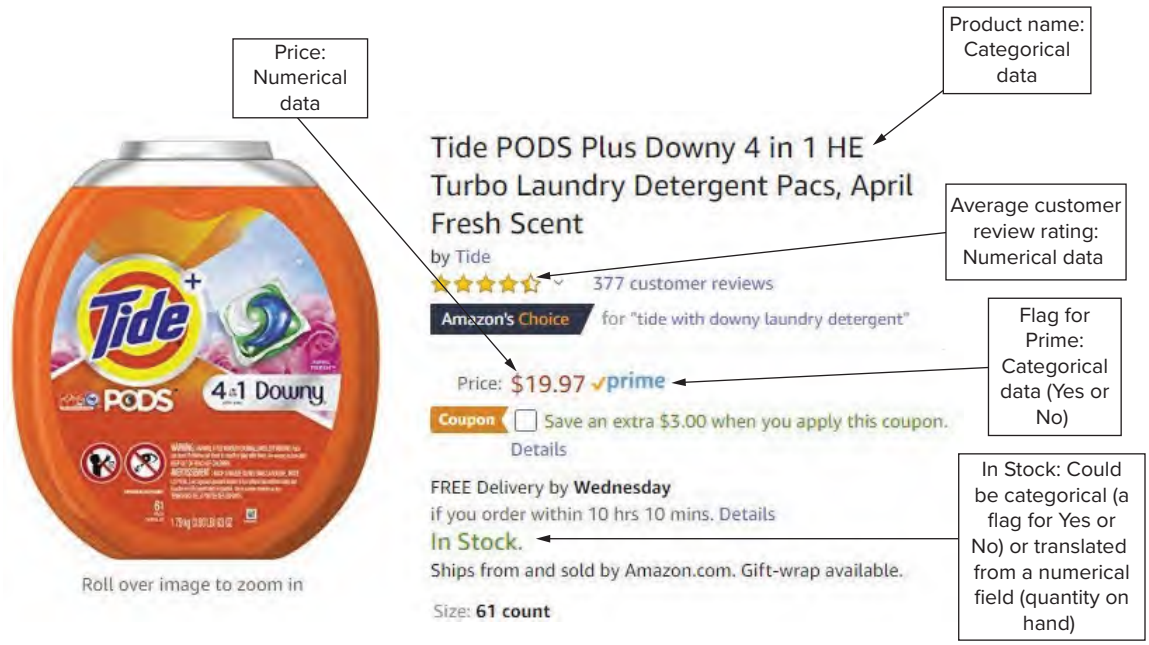


Exhibit 2.11
Tide PODS Amazon: Customer View

(Source: Amazon.com.)

Decision-Making with Data

Suppose that the makers of Tide (**Procter & Gamble**) decide to offer a $3.00 coupon to people who buy Tide Pods on Amazon. For Amazon, would the coupon data be numeric ($3.00) or categorical (a Yes/No flag to signify desire to use?)? What other data might marketers at Procter & Gamble use to promote their product? Explain whether those data are nominal, ordinal, interval, or ratio.

LAB CONNECTIONS

LABS 2.1 and **2.2** explain how to identify and work with different data types. Lab 2.1 uses Excel only, and Lab 2.2 shows alternatives using Tableau and Power BI, allowing you to compare and contrast the different software tools.

PROGRESS CHECK

13. What is the difference between interval and ratio data? Provide one example of interval data and one example of ratio data.
14. In general, how do categorical and numerical data differ? Why is it important to collect both types of data to perform your analysis?

Additional Ways to Classify Data

Databases often define data types more specifically than the four categories explained above. The following are some other common ways to categorize data.

- *String, text, short text, or alphanumeric.* A *string* of characters is a collection of one or more characters that are stored as categorical data. The characters can be letters, numbers, or a combination, but the numbers are not meaningful values and therefore cannot be used in calculations.
- *Date.* The *date* data type is a string of characters that is formatted in a traditional date format, such as mm/dd/yyyy or mm/dd/yy.
- *Number.* The *number* data type is reserved for numeric data, typically ratio data. Any characters that are stored as a number can be used in calculations.
- *Geographic.* Geographic data are any data that can be linked to a map, such as an attribute for a state, city, country, or zip/postal code. When these attributes are classified as *geographic* data, they are linked with their latitude and longitude numbers so that they can be represented on maps.
- *Dimension.* You may come across the term *dimension* in a variety of applications, including Tableau. A dimension is any attribute that is characterized as categorical.
- *Measure:* A term often used in Tableau, a *measure* is any attribute that is characterized as numerical.

As you saw in the Chapter 1 labs, a **data dictionary** is a centralized repository of information about data containing a separate record for each field/variable in the database, often detailing the data type and describing additional variables in the database.

Once you have familiarized yourself with the data types in your data set, you can turn to preparing your data for analysis. These steps are part of a common procedure called

Extract, Transform, and Load (ETL) or Extract, Load, and Transform (ELT). Even though we will focus on the "transform" step, it is important to be familiar with the term **ETL** because it is commonly used to describe the process of preparing data for analysis. As summarized in Exhibit 2.12, preparing the data requires four steps: (1) ensuring data quality, (2) validating the data for completeness and integrity, (3) cleansing the data, and (4) performing preliminary exploratory analysis.

Exhibit 2.12
The Four Steps of Data Preparation Included in ETL



LAB CONNECTION

LAB 2.4, which uses Excel, provides an example of aggregating and visualizing the different data types. Understanding how data types are determined and visualized is an important skill for business analysts.

PROGRESS CHECK

15. What is the difference between a dimension and a measure?
16. Provide an example in which a string of data that appears to be numeric could be stored as short text (that is, the string of numbers does not have mathematical meaning).

LO 2.4

Understand how to prepare data for analysis.

2.4 PREPARING DATA FOR ANALYSIS

Let's now walk through each step in preparing data for analysis.

Step 1: Ensure Data Quality

To ensure data quality, you must do the following.

1. *Make sure that the data types for each attribute are appropriate.* Sometimes items that should be stored as text are imported as numbers or vice versa. Depending on the software tool you are using to analyze the data, you may need to adjust the data type so that you can run the appropriate analysis. Ensuring data quality is particularly important in Tableau and Power BI.
2. *Check all date data types.* Many common problems revolve around the date format because a date can be presented in a variety of ways. For example, consider the different ways you can represent July 6, 2022:
   - 6-Jul-2022
   - 6.7.2022

- 44748 (When a date is formatted as a number in Excel, the date will be listed as the number of days after January 1, 1900. This feature exists to allow users to perform mathematical calculations on dates by essentially turning dates into numerical data, but it is not user friendly for interpreting the actual date itself.)
- 07/06/2022 (in the United States)
- 06/07/2022 (in Europe)

You need to format dates to match the format required by your data analysis tool and the intended recipients of your data. If the data will be used by an international audience, then it is a good idea to make your data conform to an international standard. The International Organization for Standardization (ISO) developed the ISO 8601 standard to help prevent confusion over date formats for worldwide communications. ISO 8601 indicates that you should format dates in the year-month-day format (for example, July 6, 2022 is formatted as 2022-07-06). Most professional ETL and analytical tools accept this format. If you use Excel to transform dates to this format, highlight your dates and select Home > Number > Format Cells, and choose Custom. Then type in YYYY-MM-DD and click OK.

3. *Check all numerical data types.* Numbers can be misinterpreted, particularly if they have been manually entered. For example, you may see the capital letter I instead of the number 1; the letter "O" instead of zero (0), 3 instead of E, 7 instead of seven. Watch for invalid number formats when you start sorting and analyzing your data, and be sure to correct them. Additionally, numerical artifacts such as dollar signs, commas, and parentheses are pervasive in spreadsheet data—for example $12,345.22 or (1,422.53). As you clean the data, remove any extra characters so that numbers appear in their raw form—for example, as 12345.22 or −1422.53. (That is, do not use parentheses to designate a negative number. Instead, use the negative sign.)
4. *Check all text data types, analyzing how international characters, language differences, and measures imported.*
   a. *International characters and encoding.* When you work with data that span multiple countries, it is likely that you will come across special characters, such as accent marks (á or À), umlauts (Ü), invisible computer characters (TAB, RETURN, line break, null), or special characters that are used in query and scripting languages (*, #, ", '). If your software fails to populate your data set accurately, international characters and symbols are likely to be a cause. In many cases, you can correct these problems by using a find and replace or placing the characters in quotation marks, so that the query language ignores them.
   b. *Words and measures.* Data elements may contain a variety of words and/or measures that have the same meaning. For example: cheese or fromage, ketchup or catsup, pounds or lbs, Arkansas or AR. To analyze the comparable data, you'll need to translate them into a common format by choosing one word as the standard and replacing the equivalent words. Also make sure that the measure doesn't change the meaning. The total value in U.S. dollars is not the same amount as the total value in euros. Make sure you're comparing apples to apples, or euros to euros.

Step 2: Validate the Data for Completeness and Integrity

Whenever you move data from one location to another, some of the data can get lost during the extraction. It is critical to ensure **data completeness**, which means that the data you wish to analyze were extracted fully from their original source. You must also ensure **data integrity**, which means that none of the data were manipulated or tampered with during the extraction process.

Validating the data requires you not only to have the technical skills to perform the task but also to know your data well. For example, if you know that the original data set

contains 2,000 records, it is important to ensure that your downloaded data also contains 2,000 records. To help you detect errors in data extraction and transmission, you may find it useful to use a **checksum**, which is a digit representing the sum of the correct digits in a piece of extracted data.

To validate the data after extraction, complete the following four steps.

1. *Compare the number of records that were extracted to the number of records in the source data.* This comparison will tell you whether any data were skipped or didn't extract properly due to an error or data type mismatch. This first step is critical, but it will not provide information about the data other than ensuring that the record counts match.
2. *Compare descriptive statistics for numerical fields.* Calculating the minimums, maximums, and other descriptive statistics can help you ensure that the numeric data were extracted completely. These values will also help you recognize if you should investigate for **outliers**, which are extreme values in your data set. Some outliers are helpful in business analysis because they can signal interesting transactions that may be worth digging into, but they could also represent data-entry errors. (We discuss descriptive statistics in more detail in Chapter 3.)
3. *Validate date/time fields.* As mentioned earlier, Excel can interpret dates as numerical data. Identifying the earliest and latest valid dates in a data set can be useful in terms of ensuring that you have received the appropriate slice of data.
4. *Compare string limits for text fields.* Text fields are unlikely to cause an issue if you extracted your data into Excel because Excel allows a generous maximum character number (for example, Excel 2016 allows 32,767 characters per cell). However, if you extracted your data into a software tool that more strongly limits the number of characters in a string, you will want to compare these limits to the source database's limits per field to ensure that no characters were cut off during the extraction process.

   If you find an error and the data set is not too large, you may be able to easily find the missing or erroneous data by scanning the information with your eyes. However, if the data set is large, or if the error is difficult to find, you may need to examine how the data were extracted. In particular, data that are extracted from a database are often retrieved using **Structured Query Language (SQL)**. SQL is a universal database language that can be used to create, update, and delete records and tables in relational databases. For business analysis projects, SQL is often used to retrieve a subset of data from a database. Using SQL, an analyst can combine data from one or more tables and organize the data in a way that is more intuitive for analysis. While SQL offers an excellent way to retrieve data, there is always the possibility that errors in the code will cause data-validation issues. This is why it is important to assess the SQL code itself, fix any errors in it, and re-run the extraction.

Step 3: Cleanse the Data

After validating the data, your next step is to cleanse the data as necessary to improve the data quality and subsequent analysis. The following four techniques are often used to clean data after extraction and validation.

1. *Remove headings and subtotals.* Depending on the extraction technique used and the file type of the extraction, your data may contain headings and/or subtotals that are not useful for analysis. You can prevent this problem from happening in the extraction step of the ETL process by specifying the exact data you need and the correct format.
2. *Remove leading zeroes and nonprintable characters.* Sometimes data contain leading zeroes or phantom (nonprintable) characters. **Nonprintable characters** are characters

that our eyes can't see but that the computer interprets as a part of the string. They include white spaces, page breaks, line breaks, and tabs. They can cause trouble when joining data because, although two strings may look identical to our eyes, the computer will read the nonprintable characters and will not find a match. This problem tends to crop up when numbers or dates were stored as text in the source database but need to be analyzed as numbers.

Software packages offer several ways to identify and remove nonprintable characters. **Trim functions** remove white space on either side of a cell of text and within a string of text, except for a single space between words. **Clean functions** remove the rest of the nonprintable characters but do not impact white space.

a. Excel contains functions that allow you to trim and clean data. In an Excel file, you can add an additional column next to the column you need to clean and then run either function. Or, to save time, you can run both functions at once:

Function	Purpose
=TRIM(A1)	Trims white space on either side of a cell of text and within a string of text, except for a single space between words
=CLEAN(A1)	Cleans all nonprintable characters
=TRIM(CLEAN(A1))	Combines the trim and clean functions to remove all of the unwanted white space and nonprintable characters at once

b. Power BI and Power Query in Excel both have point-and-click functions for removing nonprintable characters in the Transform tab. If you are using Power BI, follow the steps shown in Exhibit 2.13.

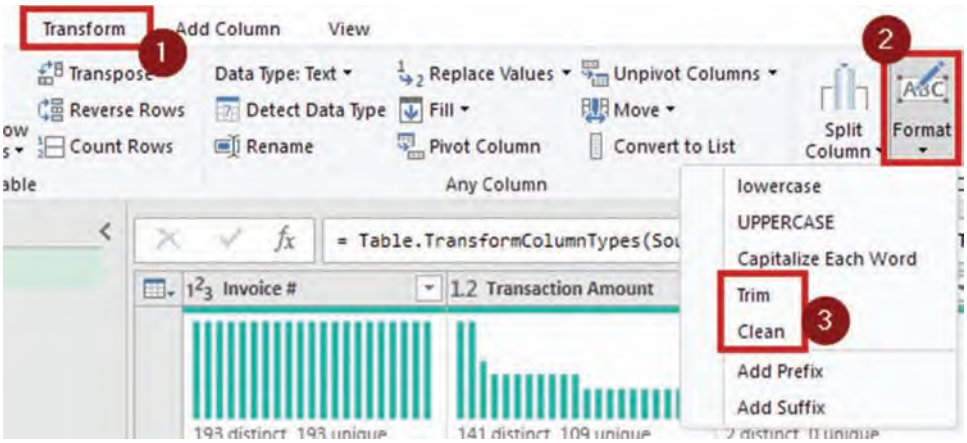


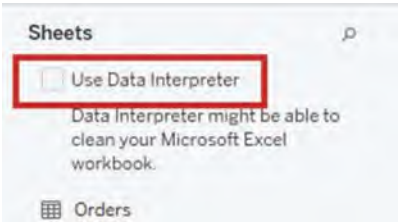
Exhibit 2.13 Transform Function in Microsoft Power BI

c. Tableau Desktop is fairly intuitive in terms of removing extra spaces and/or nonprintable characters when you connect to a data set. In addition, Tableau offers two functions that can help you cleanse the data when Tableau does not cleanse the data on its own.

- When you connect to data, place a check mark next to "Use Data Interpreter" on the Data Source page (Exhibit 2.14).

Exhibit 2.14

Tableau Software



- Once the data are loaded into Tableau, you can create a calculated field on Sheet 1 (or whichever sheet you are in) to trim any of your attributes.

Function	Purpose
LTRIM([attribute name])	Trims white space on the left side of an attribute
RTRIM([attribute name])	Trims white space on the right side of an attribute
TRIM([attribute name])	Trims white space on both sides of the attribute

Tableau does not have a function to directly clean the data (removing extra white spaces within a string of text), but you can use a combination of functions to replace spaces entirely. These functions will remove all spaces (even single spaces) and consolidate your data if you need to clean within Tableau.

Function	Purpose
REPLACE([attribute name],SPACE(1),SPACE(0))	Removes all spaces from a string of text
TRIM((REPLACE([attribute name],SPACE(1),SPACE(0)))	Trims leading and trailing spaces, as well as all spaces within the string of text

Tableau Prep is a data-preparation tool that accompanies Tableau Desktop. We do not use Tableau Prep in this textbook, but it is useful for cleansing particularly messy data sets that you need to analyze in Tableau.

3. *Format negative numbers.* If your data set includes negative numbers, you must ensure that the formatting will work for your analysis. The recommended format for negative numbers uses the negative sign instead of parentheses. Therefore, you should format negative 5.50 as −5.50, not as (5.50).
4. *Correct inconsistencies across data.* If the source database did not enforce certain rules around data entry, the data may contain inconsistencies. For example, if the data set includes a state field, Arkansas could be formatted as "AR," "Ark," or "Ar.," among other abbreviations. If you are interested in grouping data geographically, you will need to choose a common value and make all the abbreviations consistent.
5. *Decide what to do about missing values.* Missing values can occur for a variety or reasons. Perhaps some data are not applicable in a given scenario, or perhaps the necessary data were simply not collected. Regardless of the reason, missing values can lead to issues when you attempt to conduct your analysis. There are three ways to proceed; your choice will depend on the analysis you intend to perform.
   a. *Leave the missing values as they are.* If values are missing in a field because the field is not applicable to certain scenarios or because survey participants chose not to report their information (such as their phone numbers or their income level), leaving the missing values in the data can lead to meaningful analysis in understanding the audience that you are assessing.

If the missing values will prevent you from running your analysis (some predictive analytics techniques require data in every field that is input), then you have two choices:

b. *Remove the records that have missing values.* When your analysis requires all input fields to have data, the simplest solution is to remove the records that contain null values in the fields you intend to use as inputs in your analysis. However, this course of action is risky if you have to remove a significant proportion of your records. For example, if you must remove 60% or more of your records due to missing data, your analysis may be substantially affected.

c. *Impute values to replace the missing values.* When a substantial amount of data is missing, you may choose to replace (or impute) the missing fields with a representative value of the data set, such as the average. If more than 60% of values are missing, it is safer to remove those records (instead of impute values), because the imputation would end up having an outsized impact on the shape of the data. To determine the best measure to impute your missing data, it is critical that you have a good understanding of the shape of your data. There are more advanced methods for imputing data that are beyond the scope of this text.

Step 4: Perform Preliminary Exploratory Analysis

Performing preliminary exploratory analysis is an important step in reframing the initial question that you identified during the SPECIFY THE QUESTION step of the SOAR model. After you have ensured the data quality, validated the data for completeness and integrity, and cleansed the data, you can begin exploring the data to ensure that you have the data you need to answer your initial question and to determine the relevance of your question. Methods for exploring your data include assessing descriptive statistics to understand the shape of your data and how different attributes compare to others. We discuss these methods in more detail in Chapter 3. Exhibit 2.15 provides a checklist for the various tasks you have learned about in this section.

Step 1: Ensure	**Step 2: Validate**	**Step 3: Cleanse**	**Step 4: Perform**
Ensure data quality. • Ensure the data types for each attribute are appropriate. • Check all date data types for consistency. • Check all numerical attributes to ensure they are numerical. • Check all text data types for consistency and typographical errors to ensure data can be grouped properly.	Validate the data for completeness and integrity. • Compare the number of records that were extracted to the number of records in the source data. • Compare descriptive statistics for numerical fields. • Validate date/time fields. • Compare string limits for text fields.	Cleanse the data. • Remove headings and subtotals. • Remove leading zeroes and nonprintable characters. • Format negative numbers. • Correct inconsistencies across data. • Determine the best course of action for missing values (keep, delete, or replace).	Perform preliminary exploratory analysis. • Determine the relevance of the question that you identified in the SPECIFY THE QUESTION step of the SOAR model. • Ensure that you have the data you need to answer your initial question.

Exhibit 2.15

PROGRESS CHECK

17. When validating the data, it is critical to assess the data for completeness and integrity. What is the difference between data completeness and data integrity? Why are they both important?
18. What is the difference between trim functions and clean functions?
19. One item to consider when cleansing the data is what to do about missing values. Describe the three choices you have with regard to missing values.

LO 2.5

Identify common tools used to prepare data for analysis.

2.5 TOOLS USED TO PREPARE DATA FOR ANALYSIS

Many excellent software tools are available to help you prepare your data. They include the following.

- **Microsoft Excel.** Excel is the most commonly used spreadsheet software for basic data analysis. A powerful tool for exploring data, it allows you to create reports, charts, and data visualizations. It can also perform fairly advanced statistical analysis. It is available on Windows and Mac as well as through the Microsoft 365 online service (which allows simple collaboration and sharing). However, its most complete set of features and compatibility with advanced plug-ins is available only on the Windows version. One major issue with using Excel for analysis of large data sets is the limit of 1,048,576 rows due to memory constraints.
- **Tableau.** Tableau is another powerful software tool for business analysis, and it includes many of the same analysis tools provided in Excel. Tableau's biggest advantage over Excel is data visualization. Tableau will almost always default to showing the data in a visual format instead of a tabular format. These data visualizations are great for exploring data, for identifying trends and outliers, and for presenting the results when sharing the story. You can download a free student license of Tableau by visiting https://www.tableau.com/academic/students and registering with your .edu email account.
- **Microsoft Power BI.** Power BI uses simple and advanced data analytics models and visualizations that can be compiled into dashboards for easy sharing with stakeholders. It builds on data from Excel or other databases and can leverage models created with Microsoft Power Query (described below) to quickly summarize key data findings. Microsoft provides Power BI Desktop for free only on Windows or through a web-based app, but the online version does not have all of the features of the desktop version and is primarily used for sharing.
- **Microsoft Power Query.** Power Query is a tool built into Excel and Power BI Desktop on Windows. It allows Excel to connect to a variety of different data sources, such as tables in Excel and databases hosted on popular platforms such as SQL Server. Power Query makes it possible to connect, manipulate, clean, and join data so that you can pull them into your Excel sheet or use them in Power BI to create summary reports and advanced visualizations. Additionally, it tracks each step you perform so that you can apply the same transformations to new data without re-creating your work from scratch.
- **Tableau Prep.** With your free license of Tableau, you can also download a free version of Tableau Prep, which is a powerful tool for preparing data for analysis. Tableau itself has several options for transforming data once you have connected to a data source, including changing data types and creating calculated fields. Tableau Prep is a great tool to use if more extensive transformation is required before you begin working with data in Tableau. Tableau Prep allows you to easily see the distributions and summaries of every column, which can help you get to know your data and identify items that should be cleaned.
- **Alteryx.** In its ability to prepare data for analysis, Alteryx is very similar to Tableau Prep. It differentiates itself by offering more file types for the output and more advanced analytics, including predictive and prescriptive analytics.
- **Open-source tools.** Many open-source (free) software tools can be used either in conjunction with or instead of tools such as Excel and Tableau.
  - **OpenRefine** is a tool for cleaning and preparing data. With functionality similar to that of Tableau Prep and Microsoft Power Query, it is helpful as an open-source

option for users who need to clean their data but do not have access to paid software tools.

- **Gretl** is an intuitive tool for users who have a base level of statistics knowledge. You can import .csv files, .xlsx files, and files from larger systems into the tool and run a variety of statistical tests.
- **R** and **Python** are programming languages that can be used to clean data and conduct business analysis. They require more advanced training than the previously mentioned tools because they require you to learn a programming language. The other tools are more user friendly because of the graphical user interface (GUI) they provide. However, learning a programming language has big benefits. Once you know the commands you need to run to achieve your results, you will get your results much faster, especially when you're working with large data sets.

It is beyond this book's scope to work with all of these tools, so we focus on three tools: Microsoft Excel, Tableau, and Power BI.

Early in your journey to becoming a business analyst, the tools you select will depend on what is available to you and which tools you are most comfortable with. As you become more experienced and skilled, you will be able to discern when to use which tool based on the type of data cleaning, preparation, or analysis you want to perform. It is not uncommon to use multiple tools in one analysis project.

Data Analytics in Your Career

The most used and useful tool for the vast majority of business analysts is Microsoft Excel. Feeling comfortable in Excel will generally make you ready for your first internship and first analyst job. In addition, Excel serves as a great entry into using other, more sophisticated software tools.

LAB CONNECTION

LAB 2.2, which uses Tableau and Power BI, shows how to prepare each data type for analysis. As we've seen, preparation of data by data type is an important skill for business analysts.

PROGRESS CHECK

20. Explain how Microsoft Excel and Tableau are similar and how they differ.
21. Why might an analyst prefer to use an open-source tool?

ETHICS

LO 2.6

Define data ethics and describe how to gather, protect, and use personally identifiable information in an ethical manner.

2.6 GATHERING AND PROTECTING DATA ETHICALLY

Because business analysts frequently have access to data with personally identifiable information (such as customer name, birthdate, address, phone number, Social Security number, and credit card information), they must continuously ask themselves if their data gathering techniques are ethical, and if they are using the data in a responsible way that

protects individuals' identity and privacy. **Data ethics** refers to the moral responsibility associated with gathering, using, and protecting personally identifiable information.

Here are some of the questions that a company must address to handle data ethically.

Gathering Data

- Does the company send a privacy notice to individuals when their personal data are collected? Can individuals opt out of personal data collection?
- Do the company's third-party data providers follow ethical practices when gathering and sharing sensitive data?

Protecting Data

- If credit card information is taken, what assurance do customers have that their credit card number will be protected?
- Does the company keep the data secure and private, and does it have safeguards in place to protect the data?
- Has the company established effective practices to mitigate the risks of data misuse? Are penalties enforced for data misuse?

Because the ethical use of data is essential in all areas of business, this book discusses ethical issues in every chapter.

PROGRESS CHECK

22. Why should a company concern itself with the ethical data practices of their third-party data providers?
23. Why is the customer's birthdate considered to be personally identifiable information?

Chapter Summary

In this chapter, we discussed common sources for data (both internal and external) and common formats for retrieving data for analysis. We also defined and explained the two different categories of data: unstructured and structured. Unstructured data, which are generally free-form text, come from sources such as social media posts, open-response customer reviews, and call transcripts. Structured data are data that can neatly fit into a table or spreadsheet. Most of the data that we work with in this textbook are structured data provided in .xlsx or .csv format, although structured data are more commonly stored in relational databases in practice. In a relational database, data are stored across a variety of tables that are linked via primary keys (unique identifiers for every record in a table) and foreign keys (a related attribute in a different table).

Structured data can be grouped into two different data types: categorical and numerical. Categorical data do not have numerical meaning but are useful nonetheless for grouping data. For example, we can group our data by state or by customer name. Numerical data are made up of meaningful numbers that we can sum, average, multiply, or divide. In combination with categorical data, numerical data can allow for powerful comparative analysis, such as comparing average revenue by customer, by state, or by month. Categorical

data are further divided into two subcategories: nominal and ordinal. Ordinal data are more sophisticated than nominal data because of the natural ordering that occurs with ordinal data. Numerical data exist on either an interval scale or a ratio scale. The difference between the two is whether zero is meaningful—that is, whether zero implies the absence of whatever is being measured or not. When zero is meaningful, the data are measured on a ratio scale. When the data are not meaningful in those terms and zero is simply just another value on the scale, then the data are measured on an interval scale.

To prepare data for analysis, we engage in a four-step process: (1) ensure data quality, (2) validate the data for completeness and integrity, (3) cleanse the data, and (4) perform preliminary exploratory analysis. A variety of software tools exist to help you prepare your data for analysis, and several of those tools were introduced and described at the end of this chapter. In this textbook, we use mostly Excel, Tableau, and Power BI.

Finally, it is critical for analysts to consider data ethics when gathering, using, and protecting personally identifiable information.

Key Terms

aggregated data (*78*) Individual data points that have been combined into subtotals, such as counts, sums, or averages.

Big Data (*74*) Data sets that are too large and complex for businesses' existing systems to capture, store, manage, and analyze.

categorical data (*78*) Data that are categorized into groups that are represented either by words or by non-meaningful numerical data, such as grouping a collection of people by age (adult, non-adult), categorizing transaction types (sales versus returns), or specifying inventory costing method (LIFO, FIFO, average cost).

centralized database (*70*) An integrated database that contains all of a company's internal, structured data in one location.

checksum (*86*) A digit representing the sum of the correct digits in a piece of extracted data, used to detect errors in data extraction and transmission.

clean function (*87*) A computer function that cleanses data of nonprintable characters but does not impact white space.

data completeness (*85*) Occurs when data were extracted fully from their original source.

data dictionary (*83*) A centralized repository of information about data containing a separate record for each field/variable in the database, often detailing the data type and describing additional variables in the database.

data ethics (*92*) The moral responsibility associated with gathering, using, and protecting personally identifiable information.

data integrity (*85*) Occurs when none of the data were manipulated or tampered with during the extraction.

data lake (*76*) A repository for a large amount of both structured and unstructured data (internal and external) to be integrated for reporting and data analysis.

data warehouse (*76*) A repository that allows a large amount of structured data (internal and external) to be integrated for reporting and data analysis.

distributed computing (*76*) The storage of data across multiple databases, allowing greater storage of data and increased speed of analysis.

enterprise resource planning (ERP) system or **enterprise system (*70*)** Type of business management software that integrates applications from throughout the business (such as manufacturing, accounting, finance, and human resources) into one system.

ETL (Extract, Transform, and Load) (*84*) The process of preparing data for analysis; sometimes called ELT (Extract, Load, and Transform).

external data (*70*) Data that are housed outside a company; may require effort and money to access.

fields (*71*) The columns that contain descriptive characteristics about the observations in a table; also called attributes or variables.

flag (*82*) A term used to describe a data set in which there are only two options.

foreign key (*71*) A field or attribute that exists in relational databases to create a look-up relationship between two tables.

internal data (*70*) Data that are housed within a company.

Internet of Things (IoT) (*75*) A network of physical objects, such as cars and watches, that have sensors connecting them to the internet and allowing for the exchange of data.

interval data (*81*) Numerical data measured along a scale, with an equal and definitive interval or distance between each data point but no meaningful zero; examples include temperature and SAT scores.

nominal data (*79*) Categorical data that cannot be ranked. Examples include transaction type (purchase, return) and country of origin.

nonprintable characters (*86*) Characters that our eyes can't see but that the computer interprets as a part of the string. They include white spaces, page breaks, line breaks, and tabs.

numerical data (*78*) Meaningful numbers that represent quantities, such as transaction amount, age, or score on an exam.

OLAP (online analytical processing) (*76*) A computing method that enables users to easily and selectively extract and query data for analysis from different points of view.

OLAP cube (*76*) Data structure of compressed, multidimensional data for faster analytics processing.

ordinal data (*80*) Categorical data with a natural order that allows them to be ranked and sorted, such as letter grades (A, B, C, D, F) and Olympic medals (gold, silver, bronze).

outlier (*86*) An extreme value in a data set.

primary key (*71*) A field or attribute that is required to exist in each table of a relational database and serves as the unique identifier for each record in the table.

proportion (*79*) The number of observations in one category divided by the grand total of all available observations.

ranking (*80*) A position on a scale.

ratio (*82*) The numerical relationship between two variables.

ratio data (*81*) Numerical data with an equal and definitive value between each data point, and with a meaningful zero.

raw data (*78*) Data that have not been processed, cleaned, or aggregated.

records (*71*) The rows in a table, with each observation corresponding to a unique instance of what is being described in the table.

relational database (*71*) A database that compiles data into separate tables, with each table containing a unique list of the items stored; relational databases are means of storing data in order to ensure that the data are complete, not redundant, and to help enforce business rules. They also aid in communication and integration of business processes across an organization.

Securities and Exchange Commission (SEC) (*74*) U.S. government agency charged with protecting investors by requiring full and truthful financial disclosure by companies.

semi-structured data (*75*) Data with elements of both structured and unstructured data. Although semi-structured data does not have any internal structure (such as labeled columns), they may come with tags or markers explaining what the data represent.

structured data (*75*) Highly organized data that fit neatly in a table or a database. Balance sheets and income statements are examples of structured data.

Structured Query Language (SQL) (*86*) A universal database language that can be used to create, update, and delete records and tables in relational databases. For business analysis projects, SQL is often used to retrieve a subset of data from a database.

table (*71*) The presentation of data organized into sets of columns (fields) and rows (records).

tabular data (*77*) Data structured into rows and columns, with each column dedicated to one attribute of what is being described overall in the table.

trim function (*87*) A computer function that removes all white space on either side of a cell of text and within a string of text, except for single spaces between words.

unstructured data (*75*) Data without internal organization (or structure). Blogs, tweets, and pictures posted on Instagram are examples of unstructured data.

ANSWERS TO PROGRESS CHECKS

1. Internal data are housed within a company, while external data are housed outside a company. Business analysts will have more intimate knowledge of the internal data, including its qualities and its shortcomings, than they will have regarding external data. Internal data are generally easier to access than external data.
2. Companies compete against each other in the marketplace. If competitors or customers were able to access all of a company's data, then they might be able to use that information against the company. For example, if a competitor knows the secret formula for a consumer product (such as the recipe for Coca-Cola), it may be able to exploit that knowledge in the marketplace.
3. An integrated system is a single system that provides "one version of the truth," allowing people across the organization to access and work with the most current and consistent version of the data. For example, SAP is an ERP system that provides one integrated system across the entire enterprise.
4. If any employee could edit the data in an ERP system, it would be difficult to maintain the integrity of the database. Separating who has authority to edit the data from those who can only view the data is an important step in segregating duties to ensure that the ERP system does not become compromised by someone mishandling the data.
5. A primary key functions as a unique identifier in a table, while a foreign key functions as a lookup value. It is important that data in a foreign key field match data in the related table's primary key field.
6. To forecast future sales, business analysts will rely on internal data such as past sales performance, sales trends, and marketing budgets and expenses. Useful external data may include competitor data, industry data, and macroeconomic performance data, such as GDP or unemployment data.
7. Data from the U.S. Census Bureau may be helpful in terms of providing demographic information regarding potential customers (such as age, education level, or state of residence).
8. Big Data sources can improve traditional business analytics by providing information about what potential customers, competitors, and any other external stakeholders are saying about a company. For example, gathering social media data and geographic information about where potential customers are located can help business analysts spot trends that might not be apparent from internal sources of data.
9. Generic file extensions are platform-independent. Because they do not require a specific program such as Excel or Google Sheets, they are considered preferable to platform-dependent data.
10. Raw data allow business analysts to aggregate data in the way that works best for the analysis. If the data are already aggregated, then the analyst has to use the aggregation criteria decided by someone else.
11. Nominal data are categorical data that cannot be ranked, but ordinal data are categorical data that can be ranked.
12. Examples of nominal data include transaction type (sales or return). Although sales and return transactions cannot be ranked, they tell the analyst which type of transaction they are. Examples of ordinal data include dates, such as the date a transaction occurred. These dates have a natural order to them.
13. Interval data are numerical data with an equal interval or distance between each observation. Ratio data are numerical data with an equal and definitive interval between each data point and a meaningful zero. One example of interval data is how much time has passed. The distance between 1 and 2 o'clock is the same as the distance between 2 and 3 o'clock. Examples of ratio data include the sum of transactions, such as total sales and net income for a specific period.
14. Categorical data are typically text based. Even if categorical data present as a number (such as Invoice Number), the numerical value is not meaningful. Numerical data are data

that have a mathematically significant meaning, which allows aggregate measures such as sum and average to be meaningfully calculated. It is important to collect both categorical and numerical data because we can derive insights by comparing numerical data across different categories—for example, store number, customer state, and product description.

15. A dimension is a categorical variable, and a measure is a numerical variable. These terms are explicitly used in Tableau.
16. Some examples of string data that appear to be numerical are Sales Order ID, Product Number, and VIN (vehicle identification number).
17. Data completeness means that the data you wish to analyze were extracted fully from their original source, while data integrity means that none of the data were manipulated or tampered with during the extraction process. Both are important because if the data are not complete, you will not begin analyzing the data with all of the records or attributes you expected to use, and if the data do not have integrity then it does not matter if the data are complete because they are inaccurate. Decisions based on analysis from incomplete or inaccurate data will likely not be the best decisions.
18. Trim functions trim white space on either side of a cell of text and within a string of text, except for a single space between words. Clean functions remove the rest of the nonprintable characters (but do not impact white space). Trim and clean functions are often used together to remove all unwanted white space and nonprintable characters at once.
19. Three options for dealing with missing values are to (1) leave the missing values as they are, (2) delete the records that have missing values, or (3) impute values to replace the missing values. A common value to impute is the average of the field in which there are missing values. A general rule of thumb is that if more than 60% of values are missing, it is safer to remove those records (instead of impute values), because the imputation would end up having an outsized impact on the shape of the data.
20. Microsoft Excel is an industry-standard tool that helps business analysts store, analyze, and visualize data. Tableau is also a powerful tool for exploring data, but by default it presents results visually rather than in a tabular format.
21. An analyst may prefer to use an open-source tool for a variety of reasons, but one of the biggest advantages of open-source tools is that they are free, which removes the cost barrier to learning and working with a new tool. Open-source tools are also more flexible and allow analysts to customize the tool more than proprietary tools do.
22. If a company acquires data from a third-party provider, it assumes moral responsibility for the use of those data.
23. Customer birthdate, name, address, phone number, Social Security number, and credit card information are all considered to be personally identifiable information that a company might collect to facilitate its business. If these data are not protected and not used responsibly, they could fall into the hands of individuals who intend to use them unethically.



Multiple-Choice Questions

1. (LO 2.3) The text in blog posts is generally considered to be:
   a. structured data.
   b. interval data.
   c. unstructured data.
   d. ratio data.
2. (LO 2.3) Data that identify letter grades for students (A, B, C, D, F) are:
   a. ratio data.
   b. ordinal data.

c. numerical data.
d. interval data.

3. (LO 2.3) Employee salaries are examples of:
   a. interval data.
   b. ratio data.
   c. ordinal data.
   d. unstructured data.
4. (LO 2.3) If a company has 20 return transactions and 50 sales transactions, what is the proportion of return transactions?
   a. 20 ÷ 70
   b. 20 ÷ 50
   c. 70 ÷ 20
   d. 50 ÷ 20
5. (LO 2.3) If there are five transactions on 12/1, four transactions on 12/2, and nine transactions on 12/3, what proportion of the transactions that occur during the first three days of December took place on 12/1?
   a. 18 ÷ 5
   b. 5 ÷ 12
   c. 5 ÷ 18
   d. 5 ÷ 9
6. (LO 2.3) The data in the Transaction_Type column in Exhibit 2.2 are:
   a. ratio data.
   b. nominal data.
   c. numerical data.
   d. ordinal data.
7. (LO 2.3) Averaging is a summarization method for which type(s) of data?
   a. Categorical
   b. Numerical
   c. Both categorical and numerical
8. (LO 2.3) A dimension in Tableau is which type(s) of data?
   a. Categorical
   b. Numerical
   c. Both categorical and numerical
9. (LO 2.3) A measure in Tableau is which type(s) of data?
   a. Categorical
   b. Numerical
   c. Both categorical and numerical
10. (LO 2.4) Which function in Excel or Power BI can be used to remove nonprintable characters?
    a. TRIM()
    b. CLEAN()
    c. REMOVE()
    d. CUT()

Discussion Questions

1. (LO 2.4) What is ETL and how does it relate to obtaining the data?
2. (LO 2.4) Why is validating the data for completeness and integrity an important component of obtaining the data?
3. (LO 2.1) How could economic indicators, such as unemployment and GDP, help predict sales in the next quarter and the next year?
4. (LO 2.1) If you could have any internal data you wanted, which data would you use to predict **Walmart**'s sale of pecans at Thanksgiving time each year?
5. (LO 2.1) Refer to the previous question concerning the internal data that can be used to predict **Walmart**'s sale of pecans at Thanksgiving. Which external data would be helpful in predicting these sales?
6. (LO 2.3) Why is net income or net sales better defined as ratio data instead of interval data?
7. (LO 2.3) Why is temperature measured using a Fahrenheit scale not considered to be ratio data?
8. (LO 2.3) Are exam grades classified as interval data or ratio data? Why? Keep in mind that in typical exams, a score of 0 means that you answered every question wrong (or didn't answer the question at all). Does this 0 score indicate an absolute absence of knowledge? If Student A earns 40 percent on an exam and Student B earns 80 percent on the same exam, can we assume that, compared to Student A, Student B knows twice as much about the subject?
9. (LO 2.4) Which step in the ETL process (ensuring data quality, validating the data, cleansing the data, or performing preliminary exploratory analysis) do you think is the most time consuming? Which do you think is the most important? Why?
10. (LO 2.1) Why is an understanding of the way data are structured in a relational database (particularly regarding primary keys and foreign keys) important?



Brief Exercises

1. (LO 2.3) Match each data example to its appropriate data type: categorical or numerical.

Data Example	**Data Type (Categorical or Numerical)**
Fahrenheit scale	
Survey answers: agree, indifferent, disagree	
Advertising expense	
Employee choice of retirement plan vendor (Fidelity, Edward Jones, Betterment)	
Level of customer service: good, neutral, bad	
Total monthly sales	
Employee count	

2. (LO 2.3) Match each data example to its appropriate data type. Choose from structured nominal, structured ordinal, structured interval, structured ratio, or unstructured.

Data Example	Data Type (Structured Nominal, Structured Ordinal, Structured Interval, Structured Ratio, or Unstructured)
SAT scores and scores on other standardized exams	
Total assets	
Gold, silver, and bronze medals at the Olympics	
School grades (A, B, C, D, F)	
Depreciation method (declining balance, straight-line, units of production, etc.)	
Conference-call transcript	
Educational level of employees	
Credit score	
Height of WNBA basketball players	
Sales	
Hair color	

3. (LO 2.1, 2.2, 2.3, 2.4) Match each term to its definition.

Term	Definition
1. Data dictionary	a. Categorical data that cannot be ranked
2. Nominal data	b. Numerical data measured along a scale with an equal and definitive distance between each data point, but no meaningful zero
3. Ordinal data	c. Categorical data with a natural order that allows them to be ranked
4. Ratio data	d. Numerical data with an equal and definitive value between each data point, and with a meaningful zero
5. Interval data	e. A centralized repository of information about data, often detailing the data type and describing additional variables in the database. It contains a separate record for each field/variable in the database

4. (LO 2.3) Identify each summarization method as numerical, categorical, or both.

Summarization Method	Numerical, Categorical, or Both
Counting and grouping	
Averaging	
Proportion	
Ranking	
Summing	

5. (LO 2.3) What is the data type for each of these variables in a products table? Choose from short text, long text, currency, Y/N flag, or number.

Field Name	Data Type (Short Text, Long Text, Currency, Y/N Flag, Number)
Product name	
Product description	
Sales price	
Coupon available	
Free shipping available	
Count on hand	

6. (LO 2.5) Match the software tool (Alteryx, Excel, Gretl, Power BI, Python, Tableau, Tableau Prep) with its description.

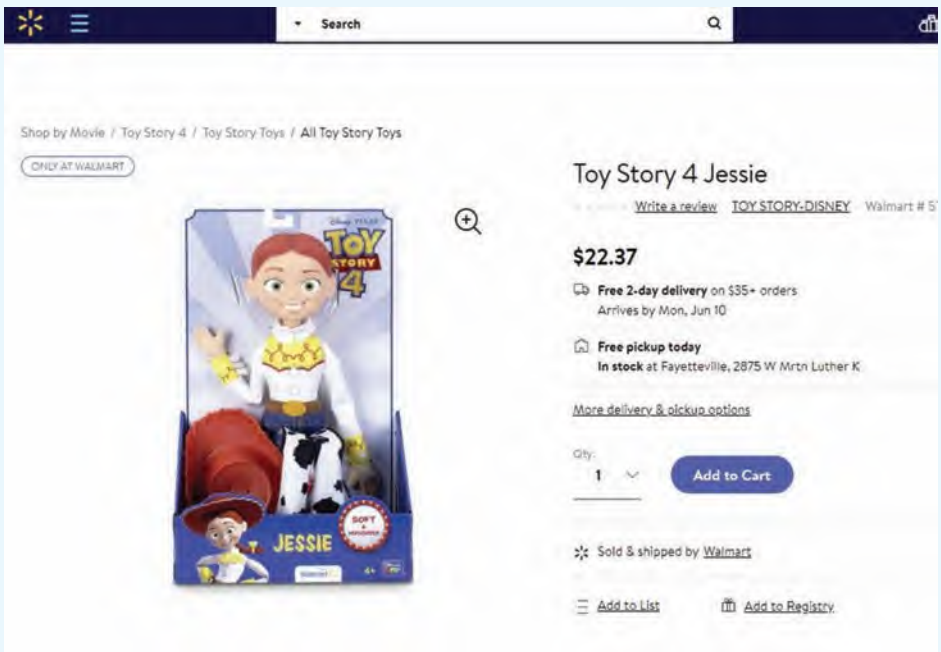
Description of Software Tool	Software Tool
Similar to Tableau Prep, but offers more file types for the output and more advanced analytics	
Usually considered the best software tool for visualizations	
Multipurpose tool for basic data preparation, business analysis, and data visualizations	
Requires knowledge of programming to use for business analysis	
Works with Tableau as a data-preparation tool	
Intuitive open-source tool for users with a base level of statistics knowledge	
Microsoft product that goes beyond Microsoft Excel for reporting and visualizing data	

Problems

1. (LO 2.3) In the Mini Case on **Tide** PODS in this chapter, you saw how a customer-facing display of data on **Amazon**'s website reflects different data types. Based on the data types identified in the Mini Case, it is easy to imagine what a row of data in Amazon's Products table might look like:

Product Name	Price	Average Customer Rating	Prime	In Stock
Tide PODS Plus Downy...	$ 19.97	4.5	Y	Y

Using the Toy Story 4 Jessie product listed on **Walmart**'s website (shown below), create a simplified data table similar to the exhibit above, including all data you deem relevant.



Walmart.com

2. (LO 2.3) Northwind Traders is a fictitious specialty foods company created by Microsoft, which has also created a fictitious database for the company. The Northwind database includes several tables, including Products, Suppliers, and Categories. Pictured below is a subset of the Products table, which details all the products that Northwind Traders sells.

ProductName	SupplierID	CategoryID	QuantityPerUnit	UnitPrice	UnitsInStock	UnitsOnOrder	ReorderLevel	Discontinued
Chai	1	1	10 boxes x 20 bags	18	39	0	10	0
Chang	1	1	24 - 12 oz bottles	19	17	40	25	0
Aniseed Syrup	1	2	12 - 550 ml bottles	10	13	70	25	0
Chef Anton's Cajun Seasoning	2	2	48 - 6 oz jars	22	53	0	0	0

   a. Review each variable in the Products table. Which would be classified as numerical data—that is, data that would be interesting to add or average?
   b. If you averaged everything in the SupplierID field, would you obtain a useful number? Explain.
   c. If you averaged UnitPrice, would you obtain a useful number? Explain.

3. (LO 2.3) Northwind Traders is a fictitious specialty foods company created by Microsoft, which has also created a fictitious database for the company. The Northwind database includes several tables, including Products, Suppliers, and Categories. Pictured below is a subset of the Products table, which details all the products that Northwind Traders sells.

ProductName	SupplierID	CategoryID	QuantityPerUnit	UnitPrice	UnitsInStock	UnitsOnOrder	ReorderLevel	Discontinued
Chai	1	1	10 boxes x 20 bags	18	39	0	10	0
Chang	1	1	24 - 12 oz bottles	19	17	40	25	0
Aniseed Syrup	1	2	12 - 550 ml bottles	10	13	70	25	0
Chef Anton's Cajun Seasoning	2	2	48 - 6 oz jars	22	53	0	0	0

   a. What does the field SupplierID represent? Is it categorical or numerical? Explain.
   b. Which two fields in this table are foreign keys? Explain.

4. (LO 2.3, 2.4) Northwind Traders is a fictitious specialty foods company created by Microsoft, which has also created a fictitious database for the company. The Northwind database includes several tables, including Products, Suppliers, and Categories. Pictured below is a subset of the Products table, which details all the products that Northwind Traders sells.

ProductName	SupplierID	CategoryID	QuantityPerUnit	UnitPrice	UnitsInStock	UnitsOnOrder	ReorderLevel	Discontinued
Chai	1	1	10 boxes x 20 bags	18	39	0	10	0
Chang	1	1	24 - 12 oz bottles	19	17	40	25	0
Aniseed Syrup	1	2	12 - 550 ml bottles	10	13	70	25	0
Chef Anton's Cajun Seasoning	2	2	48 - 6 oz jars	22	53	0	0	0

To create a meaningful report regarding when products should be reordered, which fields would you include? Which fields would you create via numerical calculations? How often would you run the report? Explain your answers.

5. (LO 2.1, 2.2) Go to **Yahoo! Finance** (https://finance.yahoo.com/) and search the symbol for **Walmart Inc.** (WMT). Click on Historical Data, and find the adjusted daily close price between February 29, 2020, and March 30, 2022.

Required

a. Take a screenshot of the first Yahoo! Finance page and label it Problem 2.5.jpg.

b. What was the adjusted close price for Walmart on March 30, 2021?

c. What was the adjusted close price for Walmart on March 30, 2022?

d. Is the adjusted closed price for Walmart considered structured or unstructured data? Explain.

e. Connecting to the data via the web is an example of connecting to external data. Explain an example of why a company would access historical stock data from this website and combine it with internal data sources for decision-making or planning.

6. (LO 2.4) The Receivables Past Due data set contains a list of outstanding invoices for a fictional company, but the data aren't ready for analysis yet. Using this data file, take the following steps to ensure data quality in Excel:

a. Assess date types. Are they uniform?

b. When date types are not uniform, analysis concerned with time series data is more difficult to perform. Adjust the data type for the Due Date so that all the data are uniform by right-clicking and selecting Format Cells.

c. The column indicating Days Past Due includes the word "days," so the data in this field cannot be used in calculations, such as computing averages and ranges of days past due. Remove the word " days" (including the leading space) from the Days Past Due field and ensure that the new data type is numerical by right-clicking and selecting Format Cells.

7. (LO 2.4) The Receivables Past Due data set contains a list of outstanding invoices for a fictional company, but the data aren't ready for analysis yet. Using this data file, take the following steps to validate and clean the data for completeness and integrity in Excel:

a. The number of records should be 200. Ensure that this number matches your data set. Explain your method for checking completeness.

b. Notice that the Customer field contains some customer names that have extraneous spaces. In column E, use the TRIM(CLEAN()) functions to clean the data in the Customer field.

d. There are several empty values in the Days Past Due field. Either remove those values or replace them with the average. Explain which method you chose and why. Also explain how you implemented the method.

e. After you have validated and cleaned the data, you can perform some preliminary exploratory analysis. Insert a pivot table:
   i. Rows: New Customer
   ii. Values: Days Past Due
   iii. Adjust the method of summarizing the values from Sum to Maximum (Max).
   iv. Explain how viewing the data in a pivot table can help you better understand the data set. How might management use this pivot table?

Data Visualization Exercises


DATA VISUALIZATION

1. (LO 2.4) The Chapter 2 Problem data set contains a list of outstanding invoices for a fictional company, but the data aren't ready for analysis yet. Load this file into Power BI and validate and clean the data. Several of these steps will be repeated from Problems 6 and 7, and it is important to learn how to make these adjustments using different tools. Once you have cleaned the data in Power BI, you will create a visualization. Take the following steps:
   a. In the Navigator window, select **Suggest Tables > Table 1** (Sheet 1).
   b. Click Transform Data to open the Power Query Editor.
   c. Notice that Power BI automatically standardized the date format in the Due Date column.
   d. Right-click the Days Past Due column and **select** Replace Values to find and replace the string " days" (include the leading space). Type days with a leading space in the Value to Find field and leave the Replace With field empty to remove the unnecessary string.
   e. Adjust the Days Past Due field to a numerical data type.
   f. Right-click the Customer field and **select** Transform to trim and clean the data. **Select** both options for Trim and Clean.
   g. Click Close & Apply to begin exploratory analysis.
   h. Place a check mark next to Customer and next to Days Past Due.
   i. In the Values field, right-click Days Past Due to adjust the summary measure to Maximum. This will display the customers with the invoices that are the most days past due.
2. (LO 2.4) This Chapter 2 Problem data set contains a list of outstanding invoices for a fictional company, but the data aren't ready for analysis yet. Load this data file into Tableau and validate and clean the data by taking the following steps:
   a. Click the "Use the Data Interpreter" checkbox. What changes occurred to the data? Why are these changes helpful?
   b. Click into Sheet 1 and create two new calculated fields:
      i. To trim and clean the Customer field, create a new calculated field named "New Customer" with the following calculation: TRIM(REPLACE([Customer],SPACE(1), SPACE(0)))
      ii. To convert the Days Past Due field to a numerical field, take the following two actions:
         1. Create a new calculated field named "Days Past Due – Numerical" with the following calculation: REPLACE([Days Past Due]," days","")
         2. Click the Abc icon next to the new field and adjust the data type to a Whole Number.
      iii. Begin exploratory analysis:
         1. Columns: New Customer
         2. Rows: Days Past Due – Numerical
            a. Right-click Days Past Due – Numerical and **select Measure > Maximum.**
      iv. How does viewing these data in a visualization aid in understanding the data and spotting trends?

CHAPTER 2 LABS

LAB 2.1 **Excel:** Identifying and Working with Different Data Types
LAB 2.2 **Tableau:** Preparing Different Data Types for Analysis
LAB 2.2 **Power BI:** Preparing Different Data Types for Analysis
LAB 2.3 **Tableau:** Conducting Preliminary Exploratory Analysis
LAB 2.3 **Power BI:** Conducting Preliminary Exploratory Analysis
LAB 2.4 **Excel:** Aggregating and Visualizing Different Data Types

Lab 2.1 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Identifying and Working with Different Data Types

Keywords

Data Types, Numerical Data, Categorical Data, Count, Average, Sum, Proportion, Rank (Sort)

Decision-Making Context

Sometimes data types can cause unintentional errors or confusion in business analysis. Taking the time to assess and set the data type in an Excel worksheet can ease your data analysis work.

Required

1. Calculate the grand totals of Customer Count, Sum of InvoiceAmount, and Count of Due Date.
2. Subtotal the accounts receivable by date in a pivot table.
3. Add in conditional formatting to the count of the Customer field.
4. Create a line chart representing customer orders by day.

Specify the Question

How do data types affect the way I can analyze data in Excel? How many payments were due on which dates?

Obtain the Data

Invoice: A list of items that were purchased (or sold) along with the amount due from the customer; also called a bill.

This lab uses the Excel file titled Lab 2.1 Data.xlsx. The data file contains 200 individual transactions representing the amount of money each customer (column A) owes (column B) and when the invoice is due (column C). All of these receivables are past their due date as of today's date, 12/31/2025.

Open Excel file Lab 2.1 Data.xlsx and browse its contents, given the data dictionary.

Data Dictionary

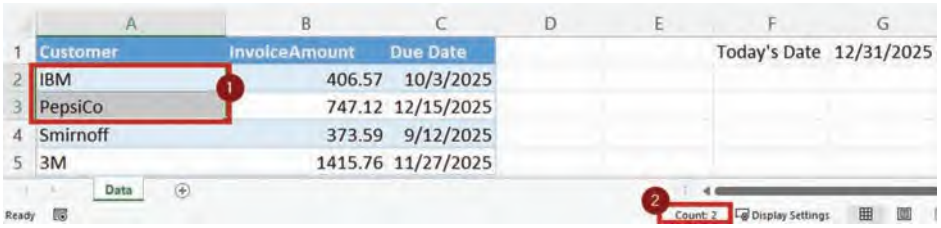
Customer:	Customer name
InvoiceAmount:	Invoice amount
Due Date:	Date that payment is due

Analyze the Data

Analysis Task 1: Default Aggregate Values in Excel's Status Bar for Different Data Types.

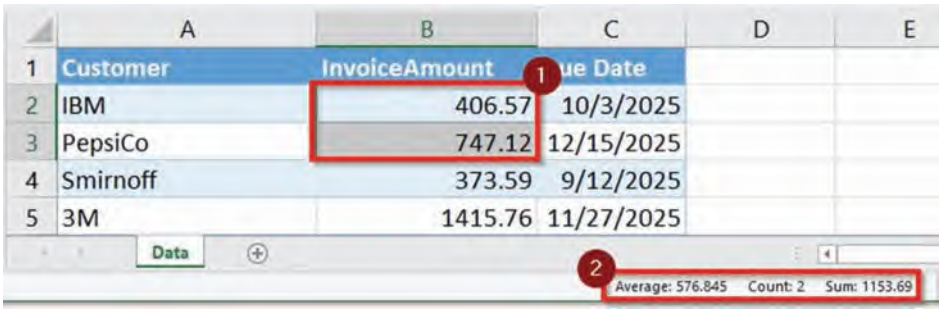
Select the first two cells in the Customer field: IBM and PepsiCo. Lab Exhibit 2.1.1 shows that a value for Count appears in the status bar on the bottom right of the Excel screen. The value for Count is 2, indicating that you have selected two items.

Lab Exhibit 2.1.1



1. Repeat the same action for the next column of data, InvoiceAmount. When you do so, you will see that not only does a value for Count appear, but so does a value for Average and Sum (as shown in Lab Exhibit 2.1.2).

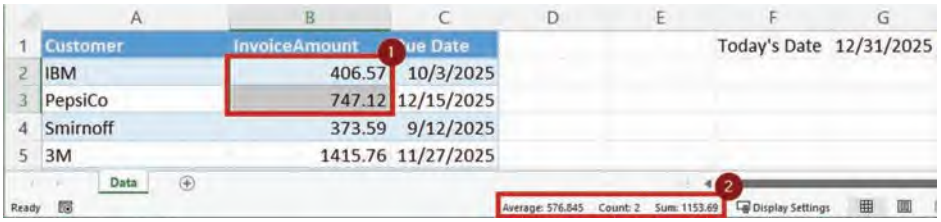
Lab Exhibit 2.1.2



Note that the data in the Customer column are *categorical,* so the only measure that can be taken is a count. The data in the InvoiceAmount column are *numerical,* so more complex mathematical measures can be taken of these data, such as sum and average.

2. Repeat the same action once again for the top two records in the Due Date column. For these two dates, you will see the numerical measures again, but they don't make immediate sense (Lab Exhibit 2.1.3).

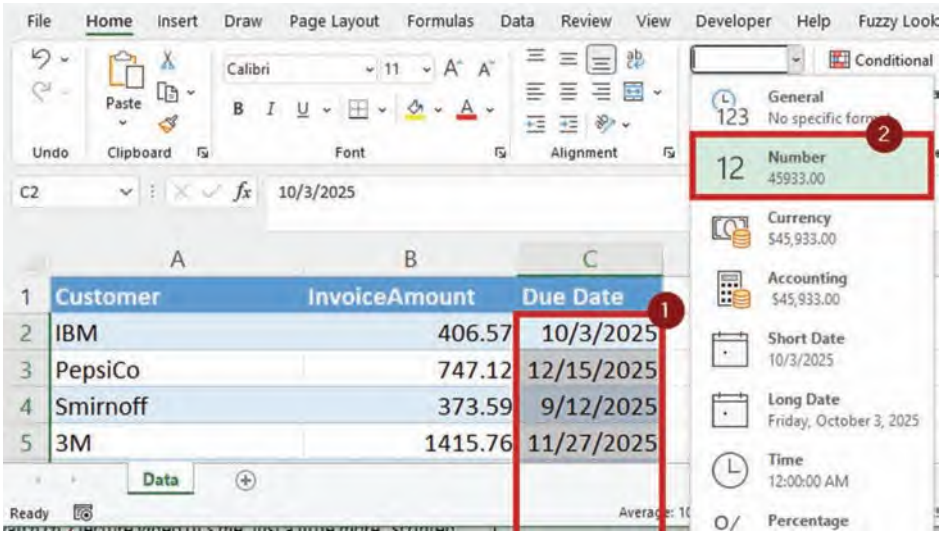
Lab Exhibit 2.1.3



When data are formatted as a Date data type in Excel, Excel associates each date with a number. The dates begin with January 1, 1900. This date is associated with the number 1 in Excel, and each day counts ascending from 1/1/1900. This great feature of Excel allows you to perform calculations on Dates; however, the math isn't always meaningful. The sum of 10/3/2022 and 12/15/2022 shown in the Status Bar as (9/18/2151) is not a meaningful value for analysis.

3. To change the data type/format for the data in the Due Date column, select all the data in that column. Then use the drop-down menu on the Home tab, Number Ribbon, to change the number format to "Number" (as shown in Lab Exhibit 2.1.4).

Lab Exhibit 2.1.4



When you make this data-type change, you will see the serial number that Excel associates with each date. Even though this is a neat change to make because it shows the underlying mechanism that Excel uses to use dates in computations, it isn't useful to view dates in this format for data analysis. It will be more meaningful to store the dates in a Date format.

4. Change the format of the Due Date column back to Short Date to restore the report's usefulness.

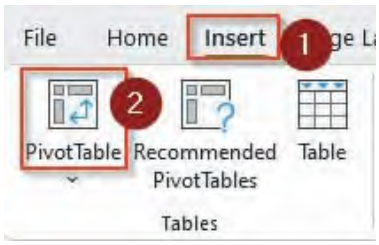
Analysis Task 2: Create More Complex Aggregations Using a Pivot Table.

Pivot Table: A tool that reorganizes and summarizes certain data using cross-tabulations without changing the underlying spreadsheet (or data).

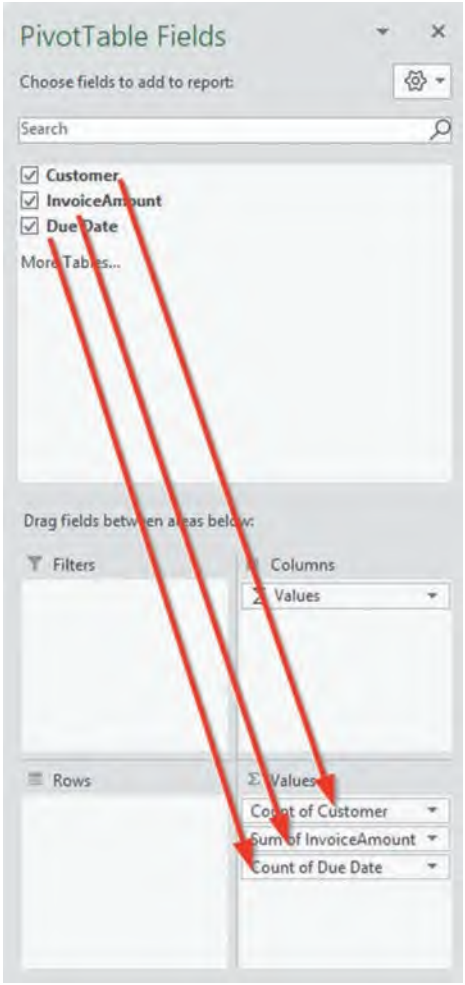
Sometimes we prefer a report with the measures that the status bar shows us. We can create such a report with pivot tables. Create a pivot table to see the measures that Excel automatically creates for each data type.

1. Ensuring that your active cell is somewhere in your table of data, create a pivot table. To do so, click PivotTable from the Insert tab (as shown in Lab Exhibit 2.1.5). Click OK.

Lab Exhibit 2.1.5



2. Drag all three fields into Values (as shown in Lab Exhibit 2.1.6). You will notice that the measures have defaulted to Count of Customer, Sum of InvoiceAmount, and Count of Due Date. Even though Due Date is stored as a numerical value in Excel, pivot tables are a little more intuitive than the Excel status bar, and Excel recognized that the Sum of all 200 Due Dates would not be a meaningful value.
3. Due Date can be more meaningful if you drag it out of the Values and into the Rows. Do so now. The report will now show the number of customer invoices due on each day. For example, only one customer invoice is due on 7/21/2022 and 7/22/2022, but two invoices are due on 7/25/2022. The report also shows the total invoice amount due on each day.



Lab Exhibit 2.1.6

Report the Results


DATA VISUALIZATION

You have analyzed the data types in this report, viewed summary statistics in Excel's status bar, and created summary statistics using Excel's pivot tables. Now, you will use two methods of visualizing the data to make the results easier to interpret.

Report the Results Task 1: Include Conditional Formatting to the Count of Customer Field.

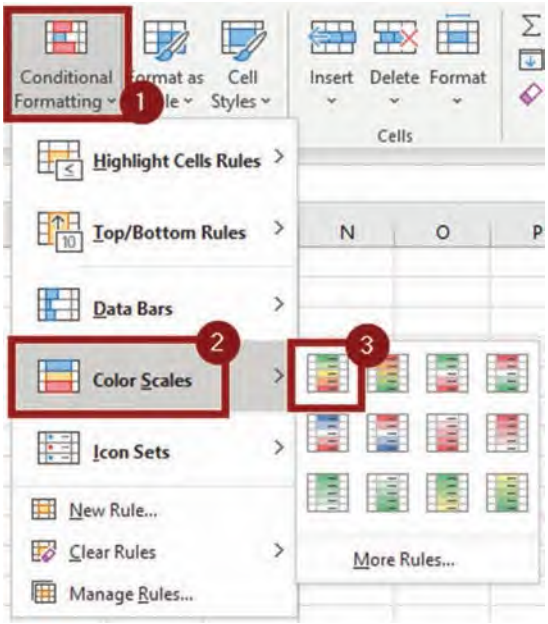
In the next two steps, you will add conditional formatting to the Count of Customer field so that it is easier to spot the extreme values and trends across the data.

1. Select all of the data in the Count of Customer field (do not include the Grand Total or either of the other columns).
2. From the Home tab in the ribbon, select **Conditional Formatting > Color Scales > Green-Yellow-Red Color Scale** (Lab Exhibit 2.1.7).

Report the Results Task 2: Include a Line Chart Representing Customer Orders by Day.

1. Remove Sum of InvoiceAmount from your pivot table. To do so, drag it out of the Values section of the Field list and drop it anywhere on the worksheet.
2. From the PivotTable Analyze tab, click PivotChart.
3. Adjust the default chart to Line and click OK.
4. Resize the Line Chart to make it easier to view, if necessary.

Lab Exhibit 2.1.7



Assessment

1. Take a screenshot of the first several rows in the pivot table with the conditional formatting, paste it into a Word document named "Lab 2.1 Excel Submission", and label the screenshot Submission 1.
2. Take a screenshot of the Line Chart, paste it into the Word document named "Lab 2.1 Excel Submission", and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 2.1 Excel Submission via Connect, if it is assigned.

Alternate Lab 2.1

Identifying and Working with Different Data Types

Apply the same steps in Lab 2.1 to the Alt Lab 2.1 Data.xlsx data set. Open Excel file Alt Lab 2.1 Data.xlsx.

Required

1. Calculate the grand totals of Customer Count, Sum of InvoiceAmount, and Count of Due Date.
2. Subtotal the accounts receivable by date in a pivot table.
3. Add in conditional formatting to the count of the Customer field.
4. Create a line chart representing customer orders by day.

Assessment

1. Take a screenshot of the first several rows in the pivot table, paste it into a Word document named "Alt Lab 2.1 Submission", and label the screenshot Submission 1.
2. Take a screenshot of the Line Chart, paste it into the Word document named "Alt Lab 2.1 Submission", and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 2.1 Submission via Connect, if it is assigned.

Lab 2.2 Tableau

Preparing Different Data Types for Analysis

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Data Types, Numerical Data, Categorical Data, Count, Average, Sum, Dimensions, Measures

Decision-Making Context

The Sláinte brewing company operates locally and also distributes its beer across several states in the United States. Specifically, it brews the following beers: Pale Ale, Imperial IPA, IPA, Imperial Stout, Stout, and Wheat. When the business analysis team at Sláinte is working with the company's sales transaction data, it's important that they recognize the data types of each attribute. When analysts do not correctly identify the data types and ensure that data have been imported appropriately, errors or confusion can occur. Taking the time to assess and set the data type as soon as you import it into Tableau can ease your data analysis work.

Required

1. Adjust the data types as necessary.
2. Aggregate the sum and then the average of all purchase orders (quantity).
3. Subtotal the purchase orders by average across raw materials.

Specify the Question

How do data types affect the way I can analyze data in Tableau?

Obtain the Data

The data set contains information about sales transactions with different customers. There are 132 unique transactions but 149 rows because some of the transactions include multiple products.

Data Dictionary

Sales Order ID:	Unique identifier for each individual sales order
Sales Order Date:	Date each sales order was placed
Sales Employee ID:	Unique identifier for the employee who was responsible for taking the sales order
Customer ID:	Unique identifier for the customer who placed the sales order
Sales Order Quantity Sold:	Quantity of each product sold in the transaction
Product Description:	Description of the product sold
Product Sale Price:	Price of each product sold in the transaction
Customer St:	State in which the customer lives
Customer City:	City in which the customer lives

1. Save the Excel file Lab 2.2 Data.xlsx to your computer; you will need to connect to the saved file from Tableau.
2. Connect Tableau to the Excel file. From Tableau, select **Connect to Data > Microsoft Excel**, then browse to the location where you saved file Lab 2.2 Data.xlsx. Then click Open.
   If the data do not load automatically, double-click on Sheet 1 to pull the data into Tableau. Notice that each of the columns from the Excel spreadsheet has imported into Tableau and that an icon appears above each column name. That icon indicates the data type associated with each column. Tableau makes a guess for each column about the type of data it holds (Lab Exhibit 2.2T.1).

Lab Exhibit 2.2T.1

Tableau Software

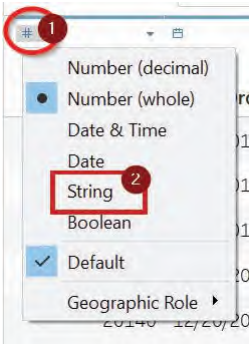
Sales Order ID	Sales Order Date	Sales Employee ID	Customer ID	Sales Order Quantity S
20001	11/1/2024	1006	2056	
20005	11/1/2024	1006	2012	
20017	11/7/2024	1006	2043	

Look at the different icons above each column in Tableau and notice that there is a number sign (#) above Sales Order ID, Sales Employee ID, Sales Order Quantity Sold, and Product Sale Price (not shown in Lab Exhibit 2.2T.1). This icon identifies the imported data as numeric data so that when you analyze the data in Tableau, you can create meaningful numerical measures such as sums or averages. The only numbers for which the data are truly numeric, though, are Sales Order Quantity Sold and Product Sale Price.

3. Click the # icon above Sales Order ID to change the data type to string (which is the same thing as text) so that Tableau will work with the data as a dimension instead of a measure (Lab Exhibit 2.2T.2).

Lab Exhibit 2.2T.2

Tableau Software



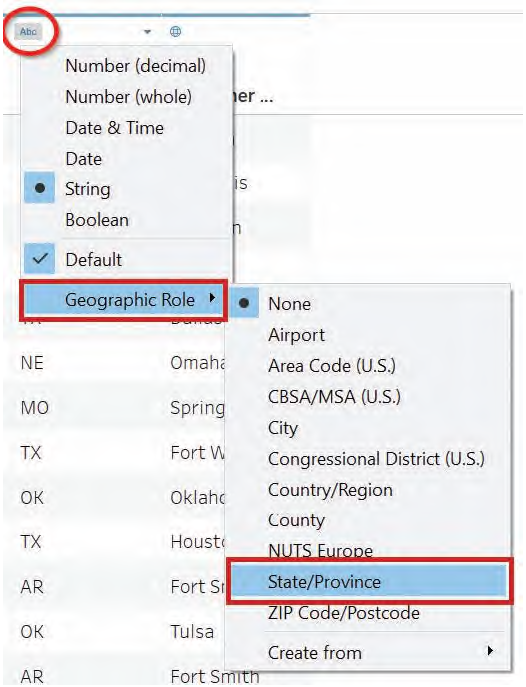


4. Make the same changes to Sales Employee ID.
5. Additionally, the data type for Customer St (state) did not import correctly. Tableau has a special data type for Geographic Data that allows you to create maps in Tableau to visualize data with geographic features. While Tableau was able to identify Customer City as a geographic data type, it likely did not do so for state. In this case, select the Abc icon above Customer St, select Geographic Role, and then select State/Province (as shown in Lab Exhibit 2.2T.3.)

Analyze the Data

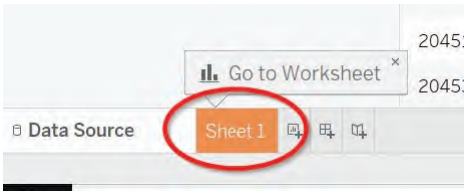
Analysis Task 1: Define Dimensions and Measures.

1. To begin working with the data, click Sheet 1 in the bottom left-hand corner (as shown in Lab Exhibit 2.2T.4.)



Lab Exhibit 2.2T.3

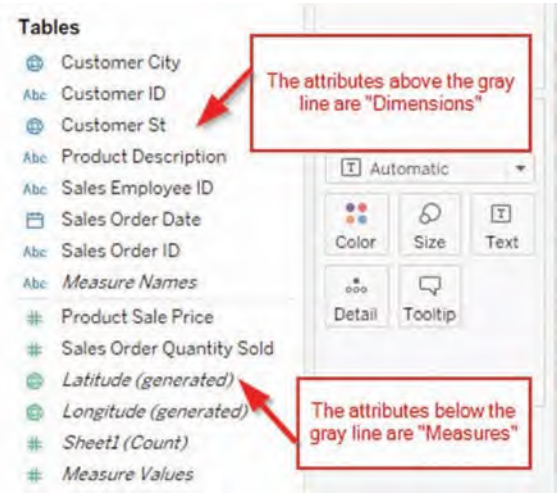
Tableau Software



Lab Exhibit 2.2T.4

Tableau Software

In Tableau, once data types are set and prepared for analysis, Dimensions and Measures are divided by a gray line, as shown in Lab Exhibit 2.2T.5. Dimensions are typically associated with categorical data and dates, while Measures are associated with numerical variables.



Lab Exhibit 2.2T.5

Tableau Software

2. To begin analyzing the data, you can double-click on any variable. Double-click on Sales Order Quantity Sold from the Measures area (as shown in Lab Exhibit 2.2T.6) to see grand total quantity sold in this data set.

Lab Exhibit 2.2T.6

Tableau Software

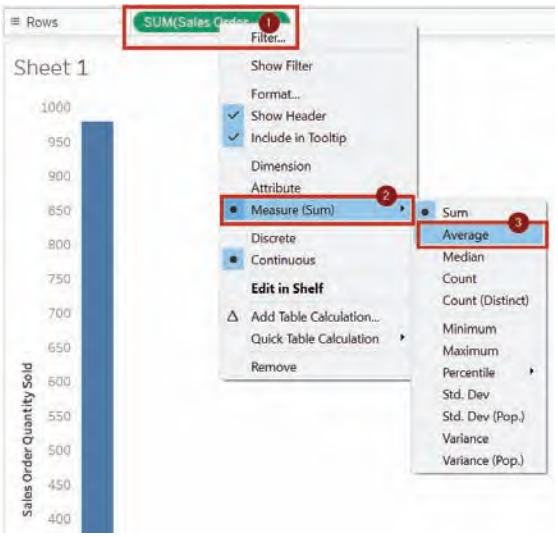


Analysis Task 2: Perform a Cross-Tabulation of Dimensions and Measures.

1. Because Sales Order Quantity Sold is a numerical measure, Tableau defaulted to aggregating the measure as a Sum. To change the default aggregate, right-click the SUM(Sales Order Quantity Sold) pill in the Rows shelf, select Measure, and then select Average (as shown in Lab Exhibit 2.2T.7).

Lab Exhibit 2.2T.7

Tableau Software



2. Enhance the meaning of this visualization by adding a categorical dimension. Double-click on Product Description from the Dimensions area to see the subtotal for each product.
3. Right-click the Sheet tab and select Rename to name the sheet Lab 2.2.

Report the Results

You have now connected to data in Tableau, transformed the variables to the appropriate data type, and begun data analysis to see how to use dimensions and measures in Tableau.

Assessment

1. Return to the Data Source in Tableau (bottom-left corner of the screen), take a screenshot of the transformed measures, paste it into a Word document named "Lab 2.2 Tableau Submission", and label the screenshot Tableau Submission 1.
2. Take a screenshot of the report output on Sheet 1 detailing the average amount of each product sold, paste it into a Word document named "Lab 2.2 Tableau Submission.doc", and label the screenshot Tableau Submission 2.
3. Answer the questions in Connect and upload your Lab 2.2 Tableau Submission.doc via Connect, if it is assigned.

Lab 2.2 Power BI

Preparing Different Data Types for Analysis

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Data Types, Numerical Data, Categorical Data, Count, Average, Sum, Dimensions, Measures

Decision-Making Context

The Sláinte brewing company operates locally and also distributes its beer across several states in the United States. Specifically, it brews the following beers: Pale Ale, Imperial IPA, IPA, Imperial Stout, Stout, and Wheat. When the business analysis team at Sláinte is working with the company's sales transaction data, it's important that they recognize the data types of each attribute. When analysts do not correctly identify the data types and ensure that data have been imported appropriately, errors or confusion can occur. Taking the time to assess and set the data type as soon as you import it into Power BI can ease your data analysis work.

Required

1. Adjust the data types as necessary.
2. Aggregate the sum and then the average of all purchase orders (quantity).
3. Subtotal the purchase orders by average across raw materials.

Specify the Question

How do data types affect the way I can analyze data in Power BI Desktop?

Obtain the Data

The data set contains information about sales transactions with different customers. There are 132 unique transactions but 149 rows because some of the transactions include multiple products.

Data Dictionary

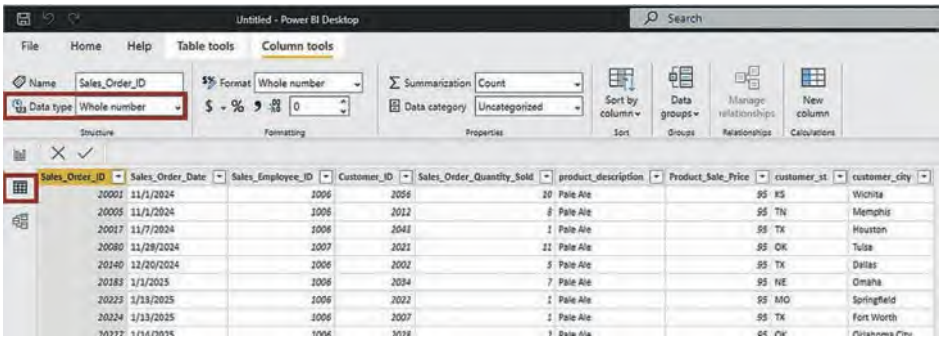
Sales Order ID:	Unique identifier for each individual sales order
Sales Order Date:	Date each sales order was placed
Sales Employee ID:	Unique identifier for the employee who was responsible for taking the sales order
Customer ID:	Unique identifier for the customer who placed the sales order
Sales Order Quantity Sold:	Quantity of each product sold in the transaction
Product Description:	Description of the product sold
Product Sale Price:	Price of each product sold in the transaction
Customer St:	State in which the customer lives
Customer City:	City in which the customer lives

1. Save the Excel file Lab 2.2 Data.xlsx to your computer. You will need to connect to the saved file from Power BI Desktop.
2. Connect Power BI to the Excel file. From Power BI Desktop, select **Home > Get Data > Excel Workbook**, then browse to the location where you have saved file Lab 2.2 Data.xlsx, and click Open. In the Navigator, check the box next to Sheet1 and click Load. Notice that each of the fields from the Excel spreadsheet appear in the Fields pane on the right side of the screen. To view and change data types, **click** the Data icon on the left side of the screen, as shown in Lab Exhibit 2.2P.1. If you click the column headers, you will see that Sales_Order_ID, Employee_ID, Sales_Order_Quantity_Sold, and Product_Sale_Price all have a data type of Whole number, indicating that the data imported as numerical data and that when you analyze the data in Power BI, you

can create meaningful numerical measures such as sums or averages. The only numbers for which the data are truly numerical, though, are Sales_Order_Quantity_Sold and Product_Sale_Price.

Lab Exhibit 2.2P.1

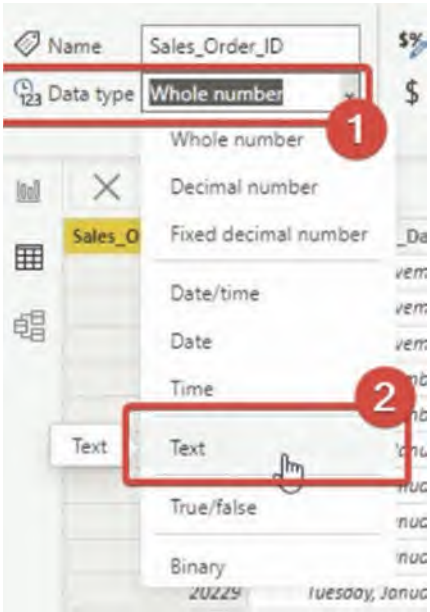
Microsoft Power BI



3. Click the Sales_Order_ID column and use the Data type drop-down menu in the ribbon to change the Data type to Text (as shown in Lab Exhibit 2.2P.2) so that Power BI will aggregate and filter the data appropriately. Perform the same action for the Sales_Employee_ID and Customer_ID fields.

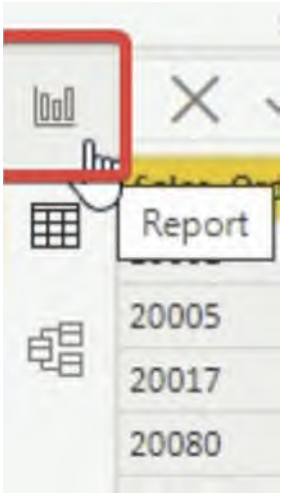
Lab Exhibit 2.2P.2

Microsoft Power BI



Analyze the Data

To begin working with the data, click the Report icon on the left (as shown in Lab Exhibit 2.2P.3).

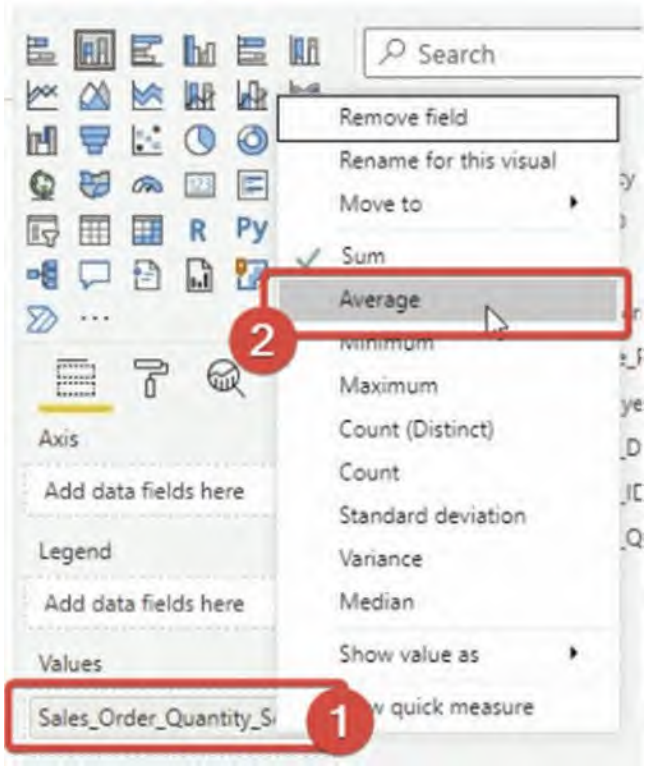


Lab Exhibit 2.2P.3

Microsoft Power BI

Analysis Task 1: Perform a Cross-Tabulation of Dimensions and Measures.

1. To begin analyzing the data, you can add a visualization for any variable. Click the Stacked Column Chart icon in the top row of the Visualizations pane (the second one from the left). Expand the Sheet1 table in Fields, and then drag Sales_Order_Quantity_Sold from the Fields pane to the Values box in the Visualizations pane to see a grand total of the quantity sold in this data set.
2. Because Sales Order Quantity Sold is a numeric measure, Power BI defaulted to aggregating the measure as a Sum. To change the default aggregate, click the down arrow next to Sales_Order_Quantity_Sold in the Values box, and then select Average (as shown in Lab Exhibit 2.2P.4).



Lab Exhibit 2.2P.4

Microsoft Power BI

3. Enhance the meaning of this visualization by adding a categorical dimension. Drag Product_Description from the Fields pane to the Axis box in the Visualizations pane to see the subtotal for each product.
4. Right-click the Sheet tab and select Rename to name the sheet Lab 2.2.

Report the Results

You have now connected to data in Power BI, transformed the variables to the appropriate data type, and begun data analysis in Power BI.

Assessment

1. Take a screenshot of the report output on Sheet 1 detailing the average amount of each product sold, paste it into a Word document named "Lab 2.2 Power BI Submission", and label the screenshot Power BI Submission 2.
2. Answer the questions in Connect and upload your Lab 2.2 Power BI Submission via Connect, if it is assigned.

Alternate Lab 2.2

Preparing Different Data Types for Analysis

Apply the same steps in Lab 2.2 to the Alt Lab 2.2 Data.xlsx data set in Tableau or Power BI. The same scenario exists as before with the Sláinte brewery, only this time you are looking at the purchasing transactions (instead of sales). The raw materials that Sláinte orders are hops, barley, malt, wheat, and seasonal items.

Required

1. Adjust the data types as necessary.
2. Aggregate the sum and then the average of all purchase orders (quantity).
3. Subtotal the purchase orders by average across raw materials.

Assessment

1. Take a screenshot of the report output on Sheet 1 detailing the average amount of each raw material purchased, paste it into a Word document named "Alt Lab 2.2 Submission", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 2.2 Submission via Connect, if it is assigned.

Lab 2.3 Tableau

Conducting Preliminary Exploratory Analysis

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Data types, Numerical Data, Categorical Data, Average, Sum, Count, Proportion, Rank (Sort)

Decision-Making Context

The Sláinte brewing company operates locally and also distributes its beer across several states in the United States. Specifically, it brews the following beers: Pale Ale, Imperial IPA, IPA, Imperial Stout, Stout, and Wheat. After the business analysis team at Sláinte has imported the relevant data into Tableau, it is important to perform some preliminary exploratory analysis. Once the data are imported in Tableau, the methods of analyzing and visualizing those data shift based on the type of data you are analyzing.

Required

1. Create the String Data sheet and work through the different ways of visualizing proportions and counts.
2. Create the Geographic Data sheet and create a filled map for quantity purchased from each state.

Specify the Question

How do data types affect the way I can analyze data in Tableau?

Obtain the Data

This lab is a continuation of Lab 2.2 Tableau, which uses the file Lab 2.2 Data.xlsx. Make sure to work through Lab 2.2 Tableau before proceeding to this lab to ensure that you have imported the data from Lab 2.2 Tableau and changed all the data types appropriately according to Lab 2.2 Tableau.

Data Dictionary

Sales Order ID:	Unique identifier for each individual sales order
Sales Order Date:	Date each sales order was placed
Sales Employee ID:	Unique identifier for the employee who was responsible for taking the sales order
Customer ID:	Unique identifier for the customer who placed the sales order
Sales Order Quantity Sold:	Quantity of each product sold in the transaction
Product Description:	Description of the product sold
Product Sale Price:	Price of each product sold in the transaction
Customer St:	State in which the customer lives
Customer City:	City in which the customer lives

Analyze the Data



DATA VISUALIZATION

Analysis Task 1: Create a Pie Chart Visualization.

Reviewing the data dictionary, you will see that there are four main data types represented in this data set: string, date, numerical, and geographic. Tableau will respond intuitively to each of these data types. The string, numerical, and geographic data types are "dimensions" in Tableau, which also means they are interpreted as categorical data.

Work through three different examples to see how you can provide context to numerical measures using text variables and number variables.

1. To begin, open (if you haven't done so already) the same Tableau workbook that you worked through in Lab 2.2 Tableau, and create a new worksheet. To do so, click the new sheet icon in the bottom left of the Tableau screen (as shown in Lab Exhibit 2.3T.1).

Lab Exhibit 2.3T.1

Tableau Software



   Remember that string (text) variables are categorical, and the two primary measures that we can calculate with string variables are count and proportion. We can create a variety of visuals for string variables, including pie charts, stacked bar charts, tree maps, bubble charts for proportion, and bar charts for comparing proportions.
2. Double-click on Product Description from the Dimensions area. Tableau will create a tabular view of each product that appears in the data set.
3. To make this output more meaningful, we can gather a count of every time each product has been purchased. This count is not the same thing as a sum because we're not considering quantity sold, but rather the number of transactions that had at least one of each product listed. Double-click Lab 2.2(Count) from the Measures area.
4. The data will present in a tabular format. To view this data set in a pie chart, expand the "Show Me" window in the top-right corner, and then click the Pie Chart option, as shown in Lab Exhibit 2.3T.2.

Lab Exhibit 2.3T.2

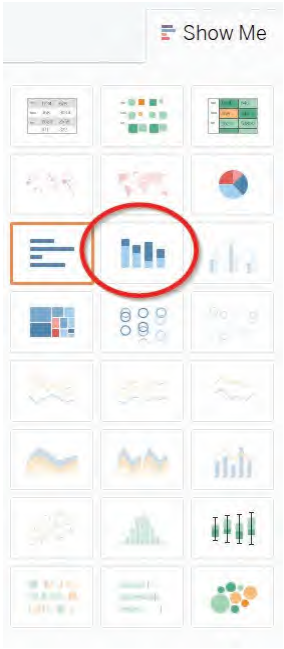
Tableau Software



Analysis Task 2: Create a Stacked Bar Chart Visualization.

To view the data in different ways, change the view to a stacked bar chart, as shown in Lab Exhibit 2.3T.3. This view will also show you a segmented view of each piece of the data to give you an idea of each product's share of the total proportion.

Pie charts and stacked bar charts are great for showing the proportion of a whole. However, if it is more important to see how the proportions of a subset of data compare against the rest (rather than seeing them as pieces of a whole), then there are a few better options.



Lab Exhibit 2.3T.3

Tableau Software

Analysis Task 3: Create a Bar Chart Visualization.

1. Continue changing your visualization by working through the following options, as noted in Lab Exhibit 2.3T.4:
   - 1. Tree map
   - 2. Packed bubble chart
   - 3. Bar chart



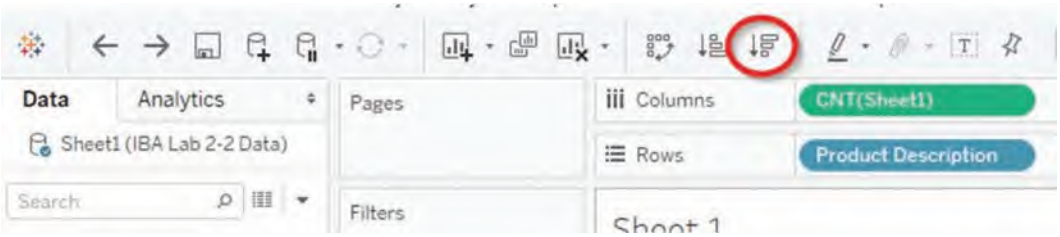
Lab Exhibit 2.3T.4

Tableau Software

Each visualization has pros and cons. It is worthwhile to run through the options whenever you are working to visualize the data to see which chart best tells the story of your analysis.

A bar chart is often the most straightforward way to tell the story because the differences between bars can be easier to see than the differences between slices of pie or shapes, particularly when the bars are sorted.

2. Once you have visualized your data in a bar chart, sort the data in a descending pattern. To do so, click the Sort Descending button (as shown in Lab Exhibit 2.3T.5).



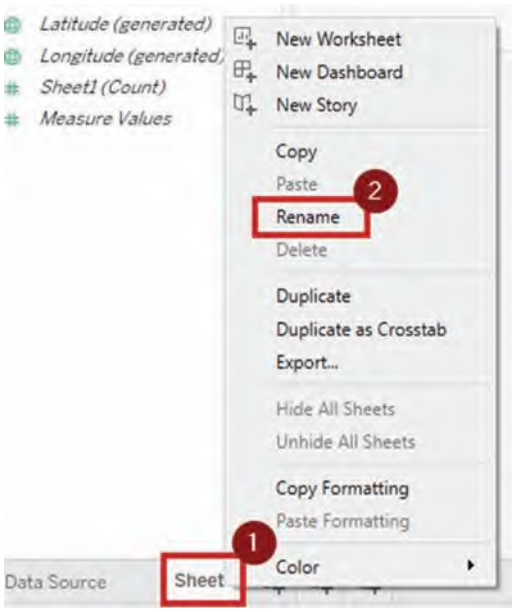
Lab Exhibit 2.3T.5

Tableau Software

3. To keep your work with the string data, rename this sheet. Right-click on the Sheet tab and click Rename. Rename the sheet to Text Data (as shown in Lab Exhibit 2.3T.6).

Lab Exhibit 2.3T.6

Tableau Software

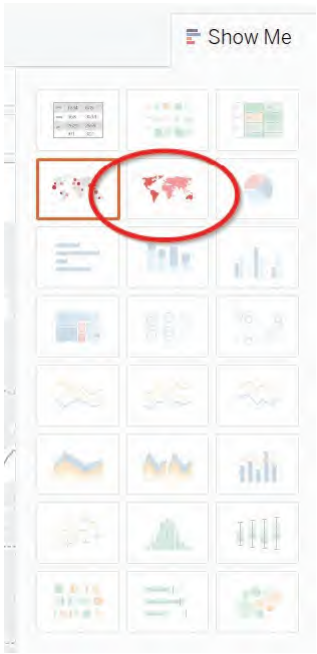


Analysis Task 4: Create a Geographic Map Visualization.

Tableau really shines with geographic variables. When variables are characterized as geographic, Tableau will default to creating a map.

1. Create a new sheet, and name it Geographic Data.
2. Double-click on Customer St. A symbol map will be automatically created, and the longitude and latitude pills will automatically appear in the columns and rows shelves, respectively. Each symbol in the map represents a state in which transactions in the data set have occurred.

3. Double-click on the measure Sales Order Quantity Sold. Notice how the symbols change size—the size of the symbol reflects the comparative amount of quantity sold in each state.
4. This map can be even easier to read if you switch it to a filled map. To do so, click Filled Map from the Show Me menu (as shown in Lab Exhibit 2.3T.7).



Lab Exhibit 2.3T.7

Tableau Software

Notice how you can immediately and easily see that most sales take place in Texas.

SOARing to Success

Data Tip

You can also visualize these data as a bar chart. Although a bar chart doesn't always make it easy to identify which region has the most sales, a bar chart might lead to quicker insights if the amount of sales between different states is quite close, therefore making it difficult to interpret the difference between the shading of different states in the filled map.

Report the Results

You have now created visualizations with different data types, compared different charts for analyzing and exploring categorical data, and created maps to analyze and explore geographic data.

Assessment

1. Return to the sheet named Text Data, take a screenshot of the bar chart, paste it into a Word document named "Lab 2.3 Tableau Submission", and label the screenshot Tableau Submission 1.

2. Return to the sheet named Geographic Data, take a screenshot of the filled map, paste it into a Word document named "Lab 2.3 Tableau Submission", and label the screenshot Tableau Submission 2.
3. Answer the questions in Connect and upload your Lab 2.3 Tableau Submission via Connect, if it is assigned.

Lab 2.3 Power BI

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Conducting Preliminary Exploratory Analysis

Keywords

Data Types, Numerical Data, Categorical Data, Average, Sum, Count, Proportion, Rank (Sort)

Decision-Making Context

The Sláinte brewing company operates locally and distributes its beer across several states in the United States. Specifically, it brews the following beers: Pale Ale, Imperial IPA, IPA, Imperial Stout, Stout, and Wheat. After the business analysis team at Sláinte has imported the relevant data into Power BI, it is important to perform some preliminary exploratory analysis. Once the data are imported in Power BI, the methods of analyzing and visualizing those data shifts based on the type of data you are analyzing.

Required

1. Create the String Data sheet and work through the different ways of visualizing proportions and counts.
2. Create the Geographic Data sheet and create a filled map for quantity purchased from each state.

Specify the Question

How do data types affect the way I can analyze data in Power BI?

Obtain the Data

This lab is a continuation of Lab 2.2 Power BI, which uses the Lab 2.2 Data.xlsx data set. Make sure to work through Lab 2.2 Power BI before proceeding to this lab to ensure that you have imported the data from Lab 2.2 Power BI and changed all the data types appropriately according to Lab 2.2 Power BI.

Data Dictionary

Sales Order ID:	Unique identifier for each individual sales order
Sales Order Date:	Date each sales order was placed
Sales Employee ID:	Unique identifier for the employee who was responsible for taking the sales order
Customer ID:	Unique identifier for the customer who placed the sales order
Sales Order Quantity Sold:	Quantity of each product sold in the transaction
Product Description:	Description of the product sold
Product Sale Price:	Price of each product sold in the transaction
Customer St:	State in which the customer lives
Customer City:	City in which the customer lives



DATA VISUALIZATION

Analyze the Data

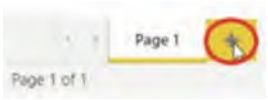
Analysis Task 1: Create a Pie Chart Visualization.

Reviewing the data dictionary, you will see that there are three main data types represented in this data set: text, date/time, and number. Power BI will respond appropriately to each

of these data types by selecting filters and aggregates that are the best for analyzing the different data types.

Work through three different examples to see how you can provide context to numerical measures using text variables and number variables.

1. To begin, open (if you haven't done so already) the same Power BI workbook that you worked through in Lab 2.2 Power BI, and create a new page by clicking the new page icon in the bottom left of the Power BI screen (as shown in Lab Exhibit 2.3P.1).

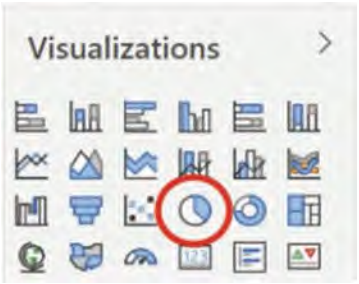


Lab Exhibit 2.3P.1

Microsoft Power BI

Remember that text variables are categorical, and the two primary measures that we can calculate with string variables are count and proportion. We can create a variety of visuals for string variables, including pie charts, stacked bar charts, tree maps, multirow cards, and waterfall charts for comparing proportions.

2. Check the box next to Product_Description in the Fields pane. Power BI will create a tabular view of each product that appears in the data set.
3. To make this output more meaningful, we can gather a count of every time each product has been purchased. This count is not the same thing as a sum because we're not considering quantity sold, but rather the number of transactions that had at least one of each product listed. Check the box next to Sales_Order_ID in the Fields pane. (Depending on your version of Power BI, you may need to change the aggregated value for Sales_Order_ID to a Count by selecting the drop-down first.)
4. The data will present in a tabular format. To view this data set in a pie chart, click the Pie Chart button in the Visualizations pane option, as shown in Lab Exhibit 2.3P.2. Then **drag** Sales_Order_ID from the Details box to the Values box in the Visualizations pane. (Depending on your version of Power BI, Sales_Order_ID may be moved to the Values box automatically after you select pie chart.)



Lab Exhibit 2.3P.2

Microsoft Power BI

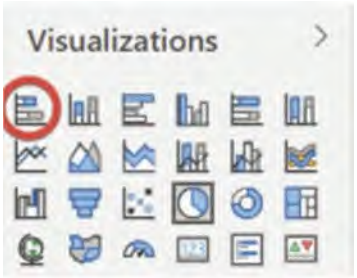
Analysis Task 2: Create a Stacked Bar Chart Visualization.

1. To view the data in different ways, change the view to a stacked bar chart, as shown in Lab Exhibit 2.3P.3. This view will also show you a segmented view of each piece of the data to give you an idea of each product's share of the total proportion.

1. Drag Product_Description from the Axis box to the Legend box in the Visualizations pane to adjust the chart as a stacked bar chart.
Pie charts and stacked bar charts are great for showing the proportion of a whole. However, if it is more important to see how the proportions of a subset of data

Lab Exhibit 2.3P.3

Microsoft Power BI



compare against the rest (rather than seeing them as pieces of a whole), then there are a few better options.

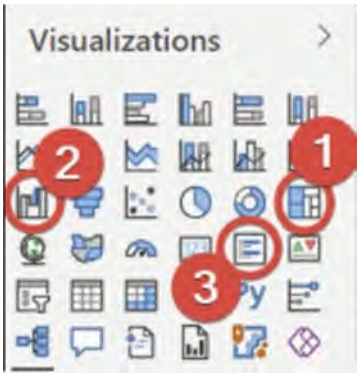
Analysis Task 3: Create a Bar Chart Visualization.

1. Continue changing your visualization by working through the following options, as noted in Lab Exhibit 2.3P.4.

- 1. Tree map
- 2. Waterfall chart
- 3. Multirow card

Lab Exhibit 2.3P.4

Microsoft Power BI



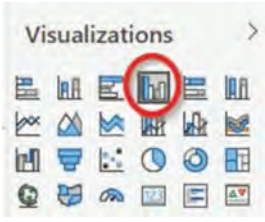
Each visualization has pros and cons. It is worthwhile to run through the options whenever you are working to visualize the data to see which chart best tells the story of your analysis.

A bar chart is often the most straightforward way to tell the story because the differences between bars can be easier to see than the differences between slices of pie or shapes, particularly when the bars are sorted.

2. Make one more adjustment to the visualization. Select Clustered Column Chart, as shown in Lab Exhibit 2.3P.5.

Lab Exhibit 2.3P.5

Microsoft Power BI

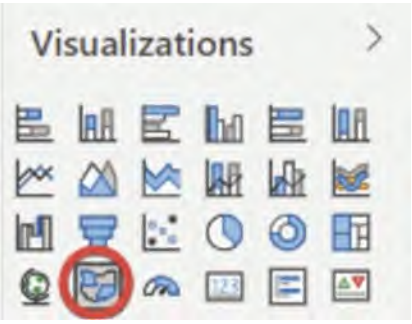


To keep your work with the text data, rename this sheet. Right-click on Page 2 and click Rename. Rename the sheet to Text Data.

Analysis Task 4: Create a Geographic Map Visualization.

Power BI can translate fields that are associated with geography (such as state) into maps. In this step, we will create a filled map representing the total quantity sold across each state.

1. Create a new page, and name it Geographic Data.
2. Click the Map button in the Visualizations pane to add a map to your page (as shown in Lab Exhibit 2.3P.6).



Lab Exhibit 2.3P.6

Microsoft Power BI

3. Drag the Customer_St field from the Fields pane to the Location box in the Visualizations pane. If this is your first time working with map data, you may receive a pop-up box indicating that map and filled map visuals are disabled. Follow the instructions on the pop-up box to enable them by going to **File > Options and settings > Options > Global > Security** and enabling the map visuals.
4. Check the box next to Sales_Order_Quantity_Sold in the Fields list.
5. To add shading and show the relative sales on your filled map, click the Format paint icon in the Visualizations tab, then expand Data colors and click the Conditional Formatting (fx) button. Adjust the Based on field to Sales_Order_Quantity_Sold (as soon as you select it, it will appear as Sum of Sales_Order_Quantity_Sold) and click OK. Notice how you can immediately and easily see that most sales take place in Texas.

SOARing to Success

Data Tip

You can also visualize these data as a bar chart. Although a bar chart doesn't always make it easy to identify which region has the most sales, a bar chart might lead to quicker insights if the amount of sales between different states is quite close, therefore making it difficult to interpret the difference between the shading of different states in the filled map.

Report the Results

You have now used Power BI to visualize a variety of different data types, compare different charts, and create a map.

Assessment

1. Return to the page named Text Data, take a screenshot of the bar chart, paste it into a Word document named "Lab 2.3 Power BI Submission", and label the screenshot Power BI Submission 1.
2. Return to the page named Geographic Data, take a screenshot of the filled map, paste it into a Word document named "Lab 2.3 Power BI Submission", and label the screenshot Power BI Submission 2.
3. Answer the questions in Connect and upload your Lab 2.3 Power BI Submission via Connect, if it is assigned.

Alternate Lab 2.3

Preliminary Exploratory Analysis

Apply the same steps in Lab 2.3 using Power BI or Tableau to the Alt Lab 2.2 Data.xlsx data set that you used in the previous lab. The same scenario exists as before with the Sláinte brewery, only this time you are looking at the purchasing transactions (instead of sales). The raw materials that Sláinte orders are hops, barley, malt, wheat, and seasonal items. You can use your Alt Lab 2.2 file as a starting point.

Required

1. Create the String Data sheet and work through the different ways of visualizing proportions and counts.
2. Create the Geographic Data sheet and create a filled map for quantity purchased from each state.

Assessment

1. Return to the sheet or page named String Data. Take a screenshot of the bar chart, paste it into a Word document named "Alt Lab 2.3 Submission", and label the screenshot Submission 1.
2. Return to the sheet or page named Geographic Data. Take a screenshot of the filled map, paste it into a Word document named "Alt Lab 2.3 Submission", and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 2.3 Submission via Connect, if it is assigned.

Lab 2.4 Excel

Aggregating and Visualizing Different Data Types

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Pivot Chart, Slicer, Pie Chart, Bar Chart

Decision-Making Context

The Sláinte brewing company operates locally and distributes its beer across several states in the United States. In particular, it brews the following beers: Pale Ale, Imperial IPA, IPA, Imperial Stout, Stout, and Wheat. After the business analysis team at Sláinte has imported the relevant data into Excel, some preliminary analysis can be performed to visualize the data.

Required

1. Create a pivot table and pivot chart to assess the sum of raw materials quantity purchased by year.
2. Create a slicer to interactively filter the pivot chart by state from which the products were ordered.
3. Adjust the pivot chart to show horizontal bars.

Specify the Question

How can we use pivot tables and pivot charts to explore different data types?

Obtain the Data

Open Excel file Lab 2.4 Data.xlsx.

The data set contains information about sales transactions with different customers. There are 132 unique transactions but 149 rows because some of the transactions included multiple products.

Data Dictionary

Attribute Name	Description	Data Type
Sales_Order_ID	Unique identifier for each individual sales order	String (text)
Sales_Order_Date	Date each sales order was placed	Date
Sales_Employee_ID	Unique identifier for the employee who was responsible for taking the sales order	String (text)
Customer_ID	Unique identifier for the customer who placed the sales order	String (text)
Sales_Order_Quantity_Sold	Quantity of each product sold on the transaction	Numerical
product_description	Description of the product sold	String (text)
Product_Sale_Price	Price of each product sold on the transaction	Numerical
customer_st	State in which the customer lives	Geographic > State
customer_city	City in which the customer lives	Geographic > City

Analyze the Data

Analysis Task 1: Create a Pivot Table.

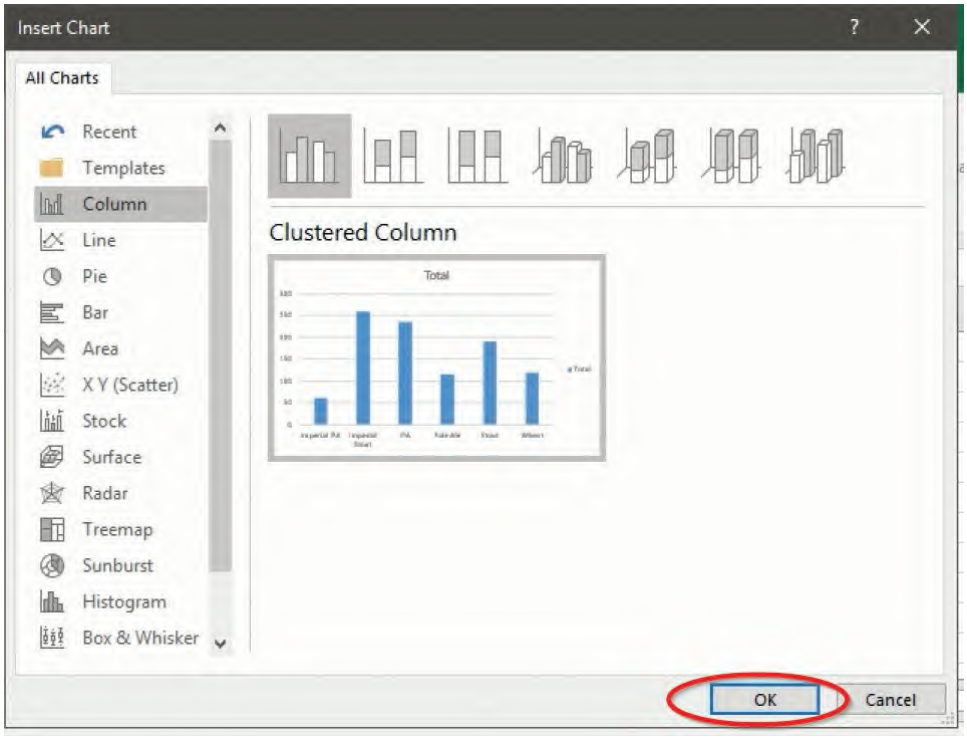
Create a pivot table that shows the quantity of each product sold.

1. From the Insert tab, click PivotTable to create a pivot table. Leave all the defaults as they are in the window that pops up and click OK.
2. Place a check mark next to Product_Description from the Field list. Notice how the data immediately default to Rows; this is how an Excel pivot table treats categorical data.
3. Place a check mark next to Sales_Order_Quantity_Sold from the Field list. Once again notice that this field defaults to Values; this is how an Excel pivot table treats numerical data.
   The pivot table should now show a tabular report of the amount sold of each product in the data set. Although this representation of the data is meaningful, the data could be more easily interpreted if they were visualized.



4. From the Analyze tab on the ribbon (ensure that your active cell is somewhere in the pivot table, which will ensure that the Analyze tab is available in the PivotTable tools tabs), select PivotChart.
5. Likely, the recommended chart for creating a pivot chart will be a Column chart. Click OK to create this chart, as shown in Lab Exhibit 2.4.1.

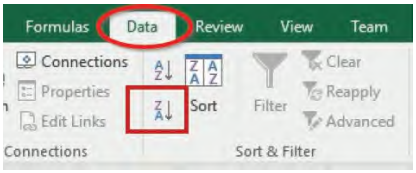
Lab Exhibit 2.4.1



This chart is meaningful but could be even easier to interpret if we reorganize the columns so that they are sorted according to the quantity of each product sold. (The default is to sort based on alphabetical order of the product description.)

6. Click one of the columns in the chart. Then, from the Data tab in the ribbon, select Z-A to sort the data, as shown in Lab Exhibit 2.4.2.

Lab Exhibit 2.4.2



Extend the pivot table to show the amount of each product sold year over year by following the item 7 directions.

7. Place a check mark next to Sales_Order_Date. If you are working with Excel 2016 (or later), you will see a date hierarchy auto-populate in the Rows section of the field list—years, quarters, months, and Sales_Order_Date. Lab Exhibit 2.4.3 illustrates how the pivot table will be collapsed to show only the year levels.

Row Labels	Sum of Sales_Order_Quantity_Sold
⊟Imperial IPA	61
2024	19
2025	42
⊟Imperial Stout	258
2024	92
2025	166

Lab Exhibit 2.4.3

Sometimes it is easier to view data related to dates when the years are in the columns instead of the rows. This ability to shift data between rows and columns is the source of the term pivot table: You can "pivot" the fields so that the table shifts depending on what variable is in the rows versus the columns.

8. Drag the "Years" field from the Rows section of the Field list into the Columns section.

SOARing to Success

Data Tip

This visual is helpful, but it is important to be wary of the months and days represented across both years. It's easy to see that 2025 has higher sales in this data set, but is the data set truly representative enough to compare quarter-to-quarter sales? It seems that we have only Quarters 1 and 2 for 2025, while we have data only for Quarter 4 in 2024. Those differences don't immediately lend themselves to year-over-year quarterly analysis. Be sure to check how closely your data collection time periods line up!

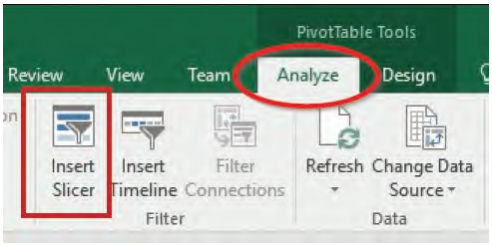
9. Remove the Quarters and Sales_Order_Date fields from the pivot table entirely. To do so, drag them into the middle of the spreadsheet and drop them outside of the Field List, which results in a pivot table by product and year.

Analysis Task 2: Add Slicers (Interactive Filters) to Explore the Data Further.

Unlike Tableau, Excel doesn't automatically create visualizations of geographic data visually with a pivot chart. You can copy and paste the data out of a pivot table into a range in Excel and create a map outside of the pivot data, but for now we will stay in the pivot table. Even without creating a map, you can still make the analysis in Excel more powerful by filtering the geographic data based on state by using Excel's slicer.

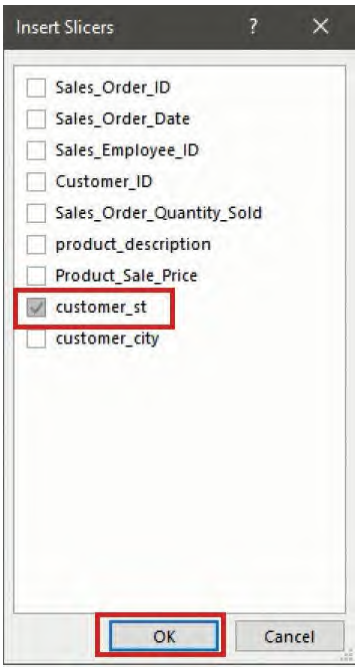
1. We can adjust our pivot table and pivot chart to view data specifically by state, rather than the data set as a whole. From the Analyze tab on the ribbon (again, ensure that your active cell is somewhere in the pivot table, which will ensure that the Analyze tab is available in the PivotTable tools tabs), click Insert Slicer, as shown in Lab Exhibit 2.4.4.

Lab Exhibit 2.4.4



2. From the Insert Slicers window that pops up, place a check mark next to customer_st and click OK, as shown in Lab Exhibit 2.4.5.

Lab Exhibit 2.4.5



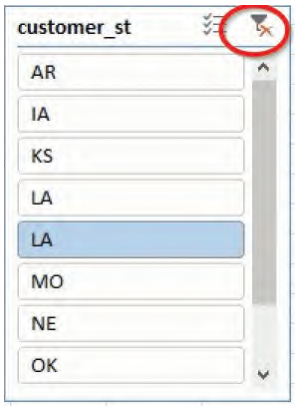
3. You have now created an interactive filter to adjust your view of quantity sold by product and by state. Click LA to see the quantity of each product sold in Louisiana.

SOARing to Success

Data Tip

Slicers are a great way to be transparent with data when you are presenting the data to others. Showing your audience exactly how you are filtering the data builds trust in you and in the data!

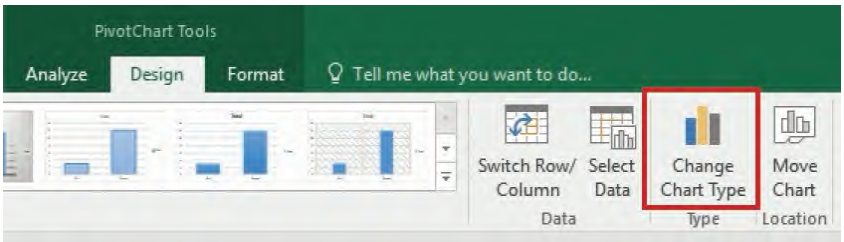
Clear the filter associated with the slicer by clicking the Clear Filter button on the Slicer window, as shown in Lab Exhibit 2.4.6.



Lab Exhibit 2.4.6

We can change the way the data are visualized by clicking Design from the PivotChart tools on the ribbon. (Ensure that your active cell is somewhere in the pivot table, which will ensure that the Design tab is available in the PivotChart tools tabs.)

4. From the Design (PivotChart tools) tab on the ribbon, click Change Chart Type (Lab Exhibit 2.4.7).



Lab Exhibit 2.4.7

5. Adjust the chart to the following chart types:

- Pie chart
- Bar chart

Share the Story

By exploring your data using pivot charts, you have learned how visualizing data makes it easier to interpret data and present the results to others.

Assessment

1. Take a screenshot of the pivot table and pivot chart without any filtering, paste it into a Word document named "Lab 2.4 Excel Submission", and label the screenshot Submission 1.
2. Take a screenshot of the pivot chart as a bar chart (horizontal bars) with the slicer in view, paste it into a Word document named "Lab 2.4 Excel Submission", and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 2.4 Excel Submission via Connect, if it is assigned.

Alternate Lab 2.4

Aggregating and Visualizing Different Data Types

Apply the same steps in Lab 2.4 to the Alt Lab 2.4 Data.xlsx data set. The same scenario exists as before with the Sláinte brewery, only this time you are looking at the purchasing transactions (instead of sales). The raw materials that Sláinte orders are hops, barley, malt, wheat, and seasonal items.

Required

1. Create a pivot table and pivot chart to assess the sum of raw materials quantity purchased by year.
2. Create a slicer to interactively filter the pivot chart by state from which the products were ordered.
3. Adjust the pivot chart to show horizontal bars.

Assessment

1. Take a screenshot of the pivot table and pivot chart without any filtering, paste it into a Word document named "Alt Lab 2.4 Submission", and label the screenshot Submission 1.
2. Take a screenshot of the pivot chart as a bar chart (horizontal bars) with the slicer in view, paste it into a Word document named "Alt Lab 2.4 Submission", and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 2.4 Submission via Connect, if it is assigned.

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Chapter 3

Analyze the Data:

Basic Statistics and Tools Required in Business Analytics

A Look Back

- To address business questions, we must know what data are available, how to access those data, and how to evaluate those data.
- Chapter 2 introduced the second step in the SOAR model: OBTAIN THE DATA. It summarized the sources of business data, the various types of data, and key data-management tools.
- Chapter 2 explained extract, transform, and load (ETL) data processes to access and prepare the data for analysis.

A Look at This Chapter

- Business analysts often have to decide between using population data versus sampling a smaller portion of the data. Chapter 3 discusses different sampling methods, along with the possibility for bias and error associated with sampling.
- This chapter also introduces the statistical toolkit required to execute the third step in the SOAR model: ANALYZE THE DATA.
- Business analysts use summary statistics and visualization methods to understand the distribution of data. They also use statistical tests, hypothesis tests, correlations, and regression to analyze data.

A Look Ahead

- Analysts perform five types of business analytics (descriptive, diagnostic, predictive, prescriptive, adaptive/autonomous), with the type of analytics depending on the question being asked.
- Chapter 4 describes descriptive analytics and diagnostic analytics, which explain what occurred and why something occurred.

Making Data-Driven Decisions

Walmart's Q2 2021 Earnings Show More Than 100% E-Commerce Growth in Last Two Years



Nadine Ghiran - SUBTA
Aug 18, 2021
4 minute read

Source: Statista, https://subta.com/walmarts-q2-2021-earnings-show-more-than-100-e-commerce-growth-in-last-two-years/. Accessed 11/9/2021.

Although **Walmart** has announced record-setting results in recent years, questions persist about its performance. Did Walmart perform greater than expectations? Is it growing faster than the retail industry? Will its performance continue in the future? How can we use visualizations and statistics to better understand what happened in the past and what is likely to happen in the future? As we learn more about business statistics and business analytics, we can better address these questions and make data-driven decisions.

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

- **LO 3.1** Distinguish between populations and samples.
- **LO 3.2** Describe sampling methods, data reduction, and types of bias in business analytics.
- **LO 3.3** Understand the basic statistics used in business analytics.
- **LO 3.4** Use software tools to create summary statistics.
- **LO 3.5** Interpret summary statistics and visualize them with histograms and box plots.
- **LO 3.6** Explain confidence intervals and hypothesis testing.
- **LO 3.7** Understand which test statistic is appropriate for the data.
- **LO 3.8** Understand correlation and regression analysis.

LABS FOR CHAPTER 3

- **LAB 3.1** **Excel:** Using Functions to Calculate Descriptive Statistics to Gain Insights About the Distribution of a Sales Data Set
- **LAB 3.2** **Excel:** Using the Excel Analysis ToolPak to Calculate Descriptive Statistics to Gain Insights About the Distribution of a Sales Data Set
- **LAB 3.2** **Tableau:** Calculating Descriptive Statistics to Gain About the Distribution of a Sales Data Set
- **LAB 3.3** **Excel:** Performing a *t*-test for Difference in Means to Determine If the Differences Between In-Person and Online Sales Are Statistically Significant
- **LAB 3.4** **Excel:** Performing an ANOVA Test for Difference in Means to Determine If There Are Significant Differences
- **LAB 3.5** **Excel:** Deriving Cost Drivers for Activity-Based Costing (Regression Analysis)

After obtaining and preparing data for analysis, you can begin the third step in the SOAR model: ANALYZE THE DATA. This chapter provides an overview of sampling methods and common statistical tests and measures that you can use to analyze data. Often when people think about data analytics, they are thinking of this step. After all, it has "analyze" in its name!

The success of your analysis depends on a well-thought-out question and appropriately prepared data. The tests you select and run in the ANALYZE THE DATA step of the SOAR model will provide the results that ultimately answer the question. Understanding the key concepts in this chapter is essential for all future chapters and for success in this course, as the different types of data analytics discussed in later chapters all make use of statistical tests and measures.

LO 3.1

Distinguish between populations and samples.

3.1 DEFINING POPULATIONS AND SAMPLES

When working with research methods and data, we need to distinguish between a population and a sample. A **population** is a group of phenomena that have something in common. The population is composed of every item of interest in that group. For example, restaurants and bricks-and-mortar retail stores often face the decision of whether to conduct business on Sundays. Some owners of retail stores and restaurants like to close on Sundays to spend time with their families, while other owners are more motivated by the additional revenue they can generate by staying open on Sundays. Suppose we want to know what percentage of restaurants and retail stores close on Sunday in the United States. To get the most accurate result, we would need to identify every restaurant and retail store in the United States. Then, we'd need to send all of them a survey (perhaps using **SurveyMonkey** or **Qualtrics**) and hope that every restaurant/retail store owner will respond. If we succeed in our goal of getting a response from every restaurant/retail store owner, then we'll have information that precisely answers our question about the entire population.

However, in reality, it is nearly impossible to get every member of the population to respond to a survey. Instead, we would analyze the responses received, which is a sample. A **sample** is simply a subset of a population, and an **inference** is a conclusion about the characteristics of a population based on the sample characteristics. Because of the challenges in reaching every member of a population, researchers and practitioners often select a subset of population members to represent that population. Their goal is to identify a sample that is representative of the total population. To continue our example, imagine that we know with certainty that 24 percent of the entire population of U.S. restaurants and retail stores is closed on Sundays. If we have chosen our sample well, then the percentage closed on Sundays should approximate 24 percent.

PROGRESS CHECK

1. What is the difference between a population and a sample? Why is it preferable to survey an entire population if doing so is possible at a reasonable cost?
2. Consider that auditors are moving from viewing only a sample of accounting transactions to viewing the full population of transactions. How will this change affect confidence in the results of the audit?

Parameters and Statistics

Two additional concepts are useful in understanding populations and samples. A **parameter** is a characteristic of a population, while a **statistic** is a characteristic of a sample. For example, using our example from above, the percentage of all U.S. restaurant/retail stores closed on Sundays might be 24 percent. This number is a parameter because it applies to the entire population.

A population parameter may or may not match the sample statistic. For example, if we survey a sample of restaurants/retail stores, the results may show that 27 percent are closed on Sundays. Here, 27 percent is a statistic. If we don't know the results for the full population, we use the sample statistic to make inferences about the population.

Organizations usually have access to some types of population data, such as the population of their current customers. When the appropriate population data already exist in your organization's database, you can use Power BI or Tableau to run analysis on very large data sets. The results will therefore be parameters rather than statistics.

Descriptive/Summary Statistics, Inferential Statistics, and Hypotheses

With this knowledge, we can further categorize statistics as either descriptive (summary) statistics or inferential statistics. **Descriptive** or **summary statistics** are measures that describe a group of interest (all of the visible components of a sample or the entire population if possible). **Inferential statistics** are measures that are calculated using only a sample of the desired population. These measures are used to test hypotheses regarding the population, and business analysts must often take extra steps after their initial calculations to be confident that the sample generalizes to the population. A **hypothesis** is a proposed explanation (or supposition) made on the basis of limited evidence as a starting point for further investigation. We will discuss hypotheses in more detail later in this chapter.

PROGRESS CHECK

3. What is the difference between a parameter and a statistic? Why might the values differ?
4. What is a hypothesis?

3.2 SAMPLING METHODS, DATA REDUCTION, AND BIAS

LO 3.2

Describe sampling methods, data reduction, and types of bias in business analytics.

Populations often have very large data sets, which can be costly and time-consuming to collect and analyze. As a solution, business analysts frequently use a subset of the data, employing sampling methods to make inferences about the population. In addition, technologies allow business analysts to reduce large data sets (perhaps data for the entire population) into smaller data sets that focus on the items of interest. Whether sampling methods or data reduction techniques are used, it is important to avoid biases.

Sampling Methods

Four common sampling methods are (1) simple random sampling, (2) stratified random sampling, (3) cluster sampling, and (4) convenience/non-probability sampling.

Simple Random Sampling

In **simple random sampling**, every observation in the population has an equal chance of being selected into the sample. This method of sampling is particularly effective if you hope to gather a representative sample of the entire population and you are not concerned about selecting a particular subset (for example, based on state, month, or another attribute) and/or when the entire population is relatively homogeneous (similar).



SOARing to Success

Using Excel to Create a Simple Random Sample

Excel can create a simple random sample in two ways. The =RAND() function generates random numbers between 0 and 1 that can be copied, sorted, and used to select the desired number of observations. The Excel Data Analysis ToolPak can also create a simple random sample.

Stratified Random Sampling

Another commonly used data-reduction method, **stratified random sampling**, entails more steps than simple random sampling. It is used when the population has clear groups (demographic or geographic, for example) that you want to ensure have adequate representation in your sample. The steps in creating a stratified sample are as follows:

1. Divide the population into groups (called *strata*).
2. Calculate the proportion of the population that each group makes up.
3. Perform a random sample on each group to ensure the appropriate number of each group (*stratum*) is represented in the overall sample.

Cluster Sampling

The technique known as **cluster sampling** begins in the same way as stratified sampling: by dividing the population into groups (called *clusters*) and calculating the proportion of the population that each group makes up. The distinction occurs in Step 3. In cluster sampling, not every group is included in the sample. This method is used when only a select few clusters are pertinent to the study, yet you still want the distribution across the groups (clusters) to be proportionate.

Convenience/Non-Probability Sampling

A method of collecting data from convenient, easy-to-access data points is **convenience sampling**, also called **non-probability sampling**. This approach is not recommended because it may not create a sample that is representative of the population, but it is used when time is limited and budgets are low. Convenience sampling takes two forms:

1. If the data are already collected, you simply select a subset of the data set.
2. If the data still need to be collected, you distribute the survey digitally or in paper format, and you stop collecting data once *n* (a particular number of) responses are received.

Data Reduction

Data reduction is the process of reducing the size of a data set to a more manageable and suitable size for a business analysis project. Technologies allow business analysts to reduce the amount of data to focus on the most critical, interesting, or abnormal items. One common method of data reduction is filtering.

Filtering

The process of choosing a smaller portion of a data set for subsequent viewing or analysis is called **filtering**. Removing the rows of data that you are not interested in analyzing, and focusing on a specific and useful subset, can speed up your analysis. Typically, you select one or more column attributes to identify the rows for removal. Excel, Tableau, and Power BI all help you filter data easily.

For example, Exhibit 3.1 shows the first several rows in a data set of sales transactions in Excel. A drop-down arrow indicates that a filter exists on a column and can be added to any column by highlighting the column and using **Data > Sort & Filter > Filter**. If you are interested in analyzing sales to customers from Texas (TX), then you can filter the data set to show only those records associated with Texas. To do so, click the drop-down button in the Customer State field (Column E), de-select Select All, select TX, and click the OK button (as shown in Exhibit 3.2).

	A	B	C	D	E	F	G	H
1	Sales Order ID	Sales Order Date	Sales Employee	Customer ID	Customer State	Product Code	Sales Order Quantity Sold	Product Sale Price
2	20003	11/1/2021	1006	2011	MO	2006	8	90
3	20005	11/1/2021	1006	2012	TN	2006	9	90
4	20001	11/1/2021	1006	2056	KS	2001	10	95
5	20009	11/2/2021	1006	2005	IA	2004	3	105
6	20012	11/4/2021	1006	2009	TX	2005	1	85
7	20013	11/4/2021	1007	2038	OK	2003	2	100
8	20020	11/7/2021	1006	2032	TX	2003	6	100
9	20017	11/7/2021	1006	2043	TX	2001	1	95

Exhibit 3.1
Example of Filtering Using Microsoft Excel

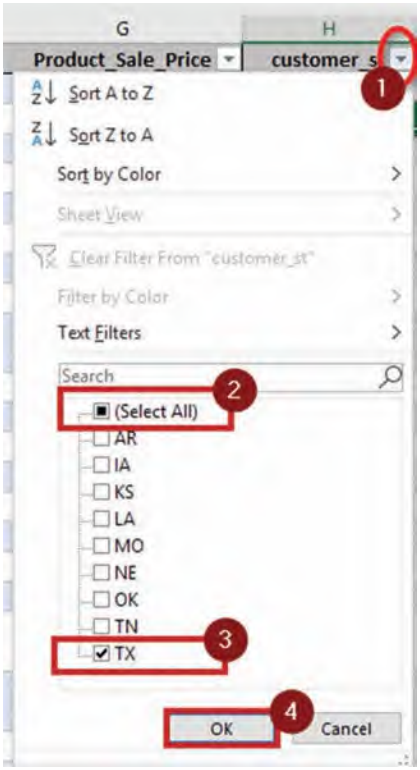


Exhibit 3.2
The Dropdown Menu Facilitates Filtering in Microsoft Excel

The Decision Process of Using a Subset of the Data

When deciding to use a subset of the data, it is important to consider the purpose of the analysis, along with the time and cost. For example, if the purpose of your analysis (based on the question) is to study a subset of data that you already have (perhaps one month of transactions or social media interaction), then you should filter the data on the items of interest. If time and money are a constraint, then you may need to use a sample of the data to speed up or manage the cost of your analysis. Running even simple calculations on a

massive data set will take much longer than it does on a subset of that data set. For example, if a company's database has 100 million rows of data that take a week to process, but an answer to a question is needed within the hour to continue operations, then analyzing a sample of the data would be the best course of action.

PROGRESS CHECK

5. Compare and contrast simple, stratified, and cluster sampling.
6. Define *convenience sampling.* Even though a convenience sample is less likely to be representative of the total population, when would it be used anyway?



Bias in Business Analytics

Bias is prejudice in favor of or against a thing, person, group, or idea. Bias can be intentional (in terms of arguing for or against something) or unintentional (resulting from poor research methods or misinterpretations of data). In terms of business analysis, bias can occur when collecting new data, conducting the analysis, and/or presenting the results. To make sure that you are engaging in ethical data analytics, you need to be aware of potential biases and work to eliminate them in the business analytics process.

Four common types of bias are (1) nonresponse bias, (2) selection bias, (3) confirmation bias, and (4) outlier bias.

Nonresponse Bias

Unless people are obligated to respond to a survey or other data-collection method, there is always a risk of nonresponse bias. Formally, **nonresponse bias** is the partiality (or bias) that results when respondents differ from non-respondents. Nonresponse bias occurs when people opt out of responding to the survey or other data-collection methods. People may not respond when the survey is poorly written, requires too much to complete, or asks personally invasive questions. Analysts can work to prevent nonresponse bias by informing survey-takers about the importance of the survey and what will be done with the results, and even by incentivizing participants. For example, you might send a $10 Starbucks gift card to people who participate, or place them in a drawing for one of five gift cards if they respond.

Selection Bias

The bias known as **selection bias** occurs when the analyst purposefully selects portions of the population that are likely to provide answers that are in line with the analyst's beliefs or hypothesis. Simple random sampling and stratified sampling are the best ways to avoid selection bias. If selection bias occurs and the results support the analyst's hypothesis, the risk is that the results do not accurately reflect the opinions or experiences of the population.

Confirmation and Outlier Bias

Two types of bias can occur after data have been collected: confirmation bias and outlier bias.

Confirmation bias occurs when analysts analyze or present results in a manner that confirms their existing beliefs or theories while ignoring data and analyses that do not confirm their beliefs. **Outlier bias** is the potential disproportionate influence of outliers in study results. *Outliers* are extreme values that substantially impact the interpretation of results. Including extreme outliers in samples can be illuminating as long as you explain the outliers, but if you do not explain outliers, then the results of your analysis may not represent the population as a whole. The result can be poor decisions.

PROGRESS CHECK

7. How can you avoid nonresponse bias in your analysis?
8. How can you avoid selection bias in your analysis?

3.3 UNDERSTANDING BASIC STATISTICS

LO 3.3

Understand the basic statistics used in business analytics.

Two important parts of analyzing data are understanding the shape of your data and calculating summary statistics. Data are captured using **random variables**, which quantify the outcomes of random occurrences. For example, a company can capture Sales Revenue by month. We can then graph a probability distribution, which shows how often different values of the random variable occur and what the distribution shape looks like. We can also calculate different statistics, including the mean, median, and mode.

Probability Distributions

A distribution of an entire population of interest allows business analysts to calculate the probability of any observation within the distribution occurring, while distributions of a sample help analysts make inferences about the population as a whole. There are several different types of data distributions, each defined by the type of data and the statistical characteristics of that data set. A **data distribution** shows all of the possible values for a variable and how often they occur or could occur.

A **probability distribution** is a statistical function that describes the possible values in a population and the likelihood that any given observation (random variable) can take a given range or value. Depending on the characteristics of the distribution, each value might have an equal chance of occurring, or some variables may be less likely to appear in the distribution. For example, Exhibit 3.3 is a probability distribution showing how long it takes a company to process (and ship) a customer's sales order. Most orders take between 7 and 12 days to process. As the bar above the 7–12 day range (or bin) shows, more than 8,000 orders were processed in this 7–12-day range. The second most frequent time is 2 to 7 days (approximately 7,500 orders). For this company, process time is most likely to be a number of days between 7 and 12.

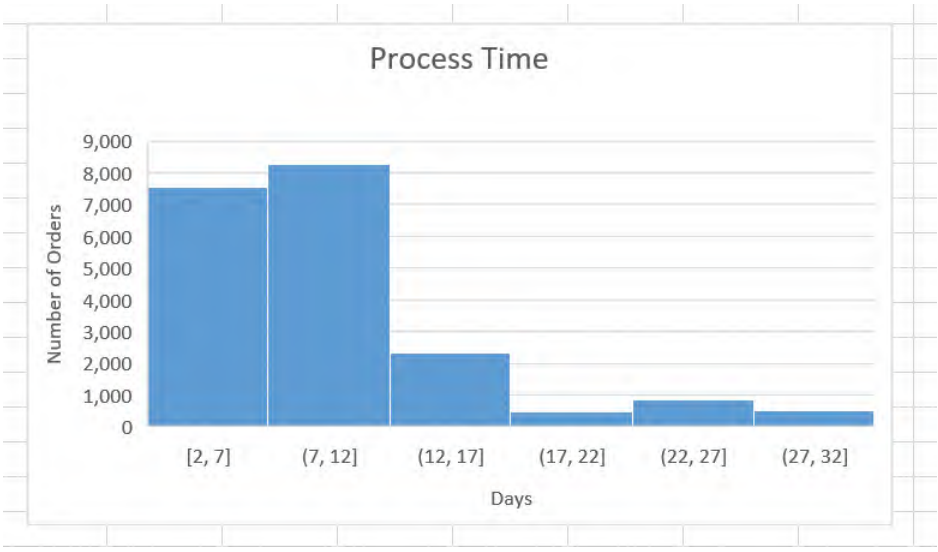


Exhibit 3.3 Example of Probability Distribution

To understand probability distributions, we must distinguish between two types of quantitative (numerical) data: continuous data and discrete data.

- **Continuous data** are numerical data that can take on any numerical value, not just whole numbers, meaning there is an infinite set of values between any two observations in a continuous data scale. Examples of continuous data are height, weight, and currency.
- **Discrete data** are numerical data that only take on whole-number (integer) values and have a finite set of values between any two observations. Examples of discrete data are number of products in inventory (you can have 0 items or 150 items in inventory, but you cannot have 147.8 items in inventory).

For any distribution, we can calculate a variety of measures to determine the shape of the data set. The type of data determines the types of probability distributions and summary measures that are appropriate. We describe these measures in the following sections, and then we explain how to derive these measures with a variety of tools: Microsoft Excel, Power BI, and Tableau. Because we focus on using software to calculate these measures, we will not dive into the raw calculations of most of these measures. For those who want more information, Appendix D provides the mathematical formulas for many of the calculations and tests mentioned in this chapter.

Measures of Central Tendency

Measures of central tendency describe the center point of a data set. They include the mean, median, and mode. These measures are important in statistics because they are the simplest way to determine the most typical point in a data set. By comparing the different measures of central tendency, an analyst can determine the shape of the distribution and better understand the concepts of symmetry and skewness.

Mean

The **mean** is the average of the measurements in a data set. For example, you can find the mean or average age of all the students in your class, the average noontime temperature over the past 30 days, or the average monthly revenue generated in a retail store over a period of 10 years. To calculate the mean, you sum all the values of a particular variable, then divide by the number of values.

The mean is the most frequently used and best-known measure of central tendency because it takes into account every value in a data set. It does have drawbacks, however. Specifically, it is susceptible to outliers. When a data set has high outlier values, then the mean will be pulled toward those higher values. When a data set has lower outlier values, then the mean will be pulled toward those lower values.

Median

Imagine arranging all of the values in your data set in order from smallest to largest. The **median** is the value that lies at the center of that arrangement. In other words, it is the midpoint of your distribution. If your data set has an even number of data points, then the median is the average of the two middle data points. The median is not susceptible to outliers. As we discuss later in this chapter, you can learn quite a bit about the shape of your data by comparing the mean to the median.

Mode

The **mode** is the most common observation in a data set. It is the simplest measure of central tendency. Imagine going back to the list of numbers you used to identify the median.

The mode is the value that appears the most often. A data distribution is *unimodal* if it has one mode, *bimodal* if it has two modes, and *multimodal* if it has more than two modes.

The mode summarizes data regardless of data type (categorical or numerical), but it is particularly important for categorical data because it is the only measure of central tendency that summarizes categorical data. (To review the different data types, see Chapter 2.) Exhibit 3.4 summarizes the calculations of the mean, median, and mode.

Exhibit 3.4
Calculating Mean, Median, and Mode

Manual Calculations of the Mean, Median, and Mode	
Data set: 1, 4, 4, 4, 5, 8, 9, 11, 14	
Mean	*Sum* all of the values, then *divide* those values by the number of values in a data set. Calculation: (1) Sum the values: 1 + 4 + 4 + 4 + 5 + 8 + 9 + 11 + 14 = 60. (2) Divide by the number of observations: 60 ÷ 9 = 6.7 Mean = 6.7
Median	*Sort* all of the values from smallest to largest in your data set, then *select* the middle value as the median. If the data set has an even number of data points, take the *average* (the mean) of the two middle data points. Calculation: (1) Arrange the values from smallest to largest: 1 4 4 4 5 8 9 11 14 ↑ (2) The median is the middle value: 5.
Mode	*Rank order* all of the values by their frequency. The value that occurs the most frequently is the mode. A data distribution is *unimodal* if it has one mode, *bimodal* if it has two modes, and *multimodal* if it has more than two modes. Calculation: The value that appears the most frequently in the data set is 4, which occurs three times. The mode is therefore 4, and this data distribution is unimodal.

Symmetry

Data are **symmetrical** if the mean, median, and mode are all equal. Symmetrical data distributions have an equal number of values on either side of the middle point of a distribution. Exhibit 3.5 shows a symmetrical distribution.

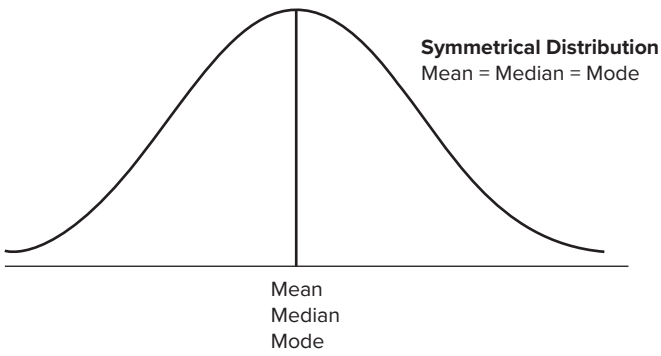
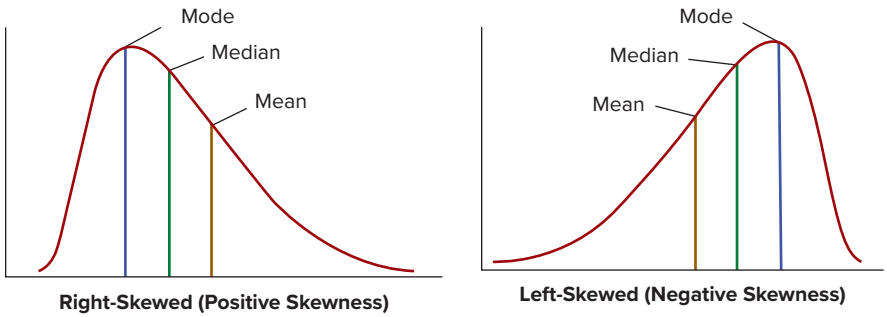


Exhibit 3.5
A Symmetrical Data Distribution

Skewness and Kurtosis

As Exhibit 3.6 shows, in some distributions the mean, median, and mode are not identical. Whenever the mean, median, and mode are not equal, the data are not only asymmetric but also skewed. **Skewed** is a more specific term than *asymmetric* because it describes

Exhibit 3.6
Right-Skewed and Left-Skewed Data Distributions



the direction of the asymmetry: Are there more values above the mean, or are there more values below the mean?

- If data are skewed to the right (positively skewed distributions), there are more observations associated with lower numbers than with higher numbers. When data are skewed to the right, or positively skewed, the mean will almost always be higher than the median because the more extreme values to the right influence the mean by pulling it higher. A naturally occurring example of a positively skewed data set is a difficult exam (where there are fewer high grades than low grades, but the high grades skew the average higher than the median).
- If data are skewed to the left (negatively skewed distributions), there are more values to the right than to the left (more large number values than small number values), so the lower values (on the left side of the distribution) tug the average to the left, making the mean is almost always lower than the median. A naturally occurring example of a negatively skewed data set is an easy exam (where there are fewer low grades than high grades, but the low grades skew the average lower than the median).

Kurtosis is another measure for describing the shape of a distribution. It describes the thickness of the tails of a distribution, specifying whether values are more clustered in the peak of a distribution or in the tails. The tails are the ends of the distribution on either side. Kurtosis and skewness work together to describe the shape of the distribution, helping analysts determine the likelihood of an event falling in a tail of a probability distribution.

Measures of Dispersion

In addition to investigating the center of your data distribution, it is also useful to understand how your data are dispersed. For example, is the majority of the data clustered tightly around the median or mean, or are the data spread out widely across the distribution? **Measures of dispersion** describe how varied, or dispersed, a data set is. They include range, interquartile range, variance, and standard deviation.

Range

The simplest measure of dispersion is the **range**, which is the difference between the maximum value and the minimum value in a data set. The range is simple to calculate, but it has limitations. Certainly a range of 150 indicates more dispersion than a range of 10, but without looking more closely at the data we cannot know if the comparatively wide range of 150 is truly indicative of how spread out all of the data points are, or whether the range is being affected by one or more outliers. We can improve on the range by using measures that take into account not just the minimum and maximum values, but rather every value in the data set.

The range has a further limitation when it is used with sample data. Specifically, it can be a poor predictor of the dispersion of data for the entire population. Two observations in a sample—even if they are the minimum and maximum values—might not necessarily reflect the population as a whole.

Interquartile Range

Another way to analyze the shape of a data set is to use the **interquartile range**, which focuses on the median. Taking the term apart a bit can help us better understand it. A *quartile* is one of four groups into which the data in a data set are grouped, with each quartile representing 25 percent, or one-quarter, of the observations. To create quartiles, we place the data in order from the lowest value to the highest value, just as we do when we want to find the median or range. The cutoff value for each quartile is a **percentile**, a value on a scale of 100 that indicates the percent of a distribution that is equal to or below it. Then we group the values as follows:

- Quartile 1: the lowest 25 percent of observations (representing all data points from the minimum value through the 25^{th} percentile).
- Quartile 2: the next 25 percent of observations; its cutoff is the median (representing all data points between the 25^{th} percentile and the 50^{th} percentile).
- Quartile 3: begins at the median and extends to the third 25 percent of observations (representing all data points between the 50^{th} percentile and the 75^{th} percentile).
- Quartile 4: the highest 25 percent of observations (representing all data points from the 75^{th} percentile to the maximum value).

Excel or Tableau ranks the observations automatically when calculating quartiles.

The prefix *inter-* in the word *interquartile* implies that we are not interested in all four quartiles but rather the middle section, specifically Quartile 2 and Quartile 3. The interquartile range is thus the range of data that surrounds the median. When you are analyzing a sample instead of a population, the interquartile range is considered more helpful than the range as a measure of dispersion because it focuses on the most common values and minimizes the risk of outliers too heavily influencing the range.

Variance and Standard Deviation

More sophisticated measures to describe the dispersion of your data in relation to the mean values are variance and standard deviation.

The **variance** is calculated by taking the average of the squared deviations from the mean for each observation, and the **standard deviation** is the square root of the variance. The standard deviation is expressed in the same unit as the data values, which makes it easier to interpret. Variance is expressed in squared values, which makes it more challenging to interpret. It's true that these terms feel very technical. Fortunately, tools such as Excel, Tableau, and Power BI help us find the variance and standard deviation without having to perform the calculations manually.

When comparing two data sets, the data set with the larger standard deviation has more variation, or dispersion, in the data around its mean. For example, financial analysts often measure the volatility of a stock price using standard deviation as a measure of risk. The greater the standard deviation (and thus, the greater dispersion around its mean), the riskier the stock is considered.

The calculations for standard deviation and for variance differ based on whether the data set consists of a sample or a population. Because we can never be 100 percent confident that our samples perfectly reflect the parameters of a population, the calculations for standard deviation for a sample differ from those for a population to minimize nonresponse bias and selection bias.

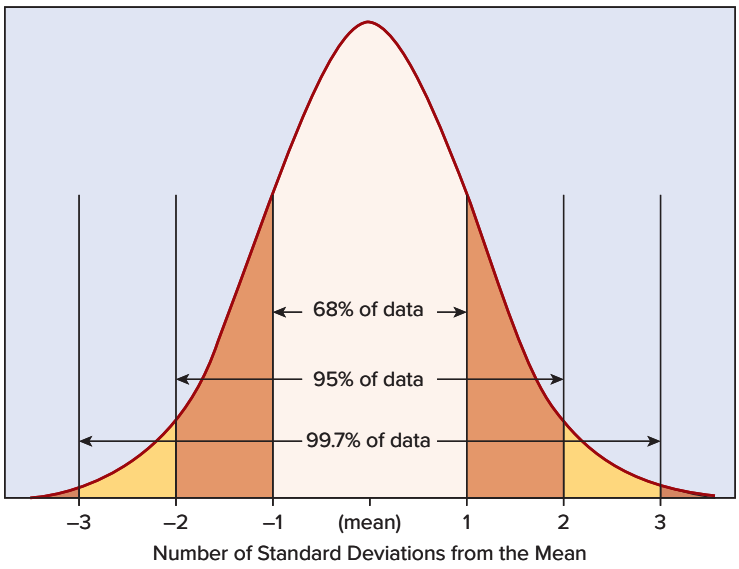
Continuous Probability Distributions

Now that we understand measures of central tendency and dispersion, let's look at some continuous data distributions that are frequently used.

Normal Distribution

The **normal (Gaussian) distribution** is a bell-shaped probability distribution that is symmetric about its mean, with the data points closer to its mean more frequent than the data points further from its mean. As Exhibit 3.7 shows, observations within one standard deviation of the mean include 68% of the data points. Observations within two standard deviations of the mean include 95% of the data points, and observations within three standard deviations of the mean include 99.7% of the data points.

Exhibit 3.7 The Normal Distribution



The normal distribution has a kurtosis of 3, so in combination with the skewness (normal distribution has a skew level of 0), the two values can help you determine how well your data distribution approximates the normal distribution.

Many of the statistical tests that we use in data analytics are based on the normal distribution and how many standard deviations a sample observation is from its mean. Many naturally occurring continuous distributions conform to the normal distribution, including newborn weight, population height, population IQ, shoe size, and employee performance.

The Standard Normal Distribution and *z*-Scores

A special case of the normal distribution is the **standard normal distribution**, which has a mean, median, and mode of 0 and a standard deviation of 1. The standard normal distribution is a theoretical distribution. That is, it does not represent a real distribution. Instead, it is used as the basis for standardizing data across different distributions to make comparisons easier or to calculate probabilities of individual observations occurring in a population.

In any distribution that we can assume to be normal, we can transform the observations into **z-scores** that represent standardized values. A *z*-score tells us how many standard deviations a data point (or observation) is from the mean. A *z*-score of (+)1 means that

the observation is one standard deviation above the mean. A z-score of –2 means that the observation is two standard deviations below the mean. For example, the mean IQ is 100 with a standard deviation of 15. Albert Einstein's IQ was reported to be 160, or four standard deviations above the mean, giving his IQ a z-score of 4 and suggesting atypically high intelligence.

Many of the statistical tests employed in data analytics are based on the standard normal distribution and how many standard deviations a sample observation is from its mean.

Uniform Distribution

A **uniform distribution** of continuous data is recognizable by its rectangular shape. It is characterized by a range of values that are all equally likely to occur. For example, if all customers are equally likely to spend between 5 minutes and 45 minutes shopping in a store, data regarding customer time in the store will roughly follow a uniform distribution, similar to the one shown in Exhibit 3.8.

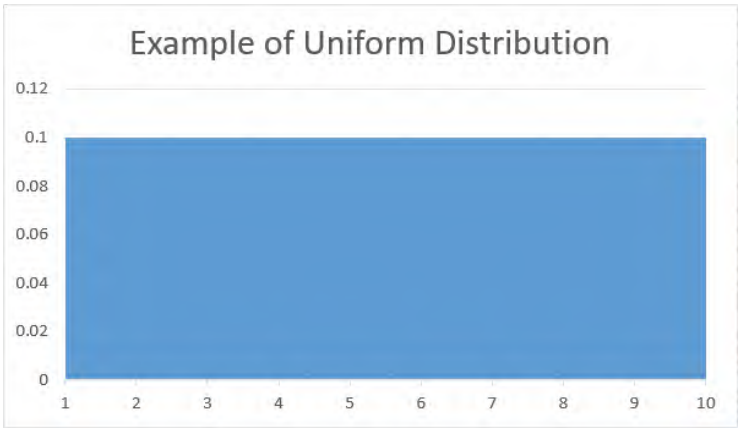


Exhibit 3.8
Example Uniform Distribution

DATA ANALYTICS AND YOUR CAREER

A deep knowledge of statistics continues to offer excellent career opportunities. According to the U.S. Bureau of Labor Statistics (BLS), 44,800 people were employed as mathematicians or statisticians in 2020, and this number is expected to grow 33 percent by 2030. This rate is much faster than the average for all occupations in the United States. In 2020, the median income for mathematicians or statisticians was $93,290.[1] An understanding of data and statistics continues to be an important and growing area in the economy.

PROGRESS CHECK

9. Why is it useful to understand the center of your data's distribution? Why is it helpful to understand how dispersed your data are?
10. What does a data set with zero skewness represent? What does a negative skewness represent? A positive skewness?
11. When comparing data sets, what purpose does the standard deviation serve? If Data Set A has a higher standard deviation than Data Set B, what does that tell you?

[1]U.S. Bureau of Labor Statistics, May 2020, https://www.bls.gov/ooh/math/mathematicians-and-statisticians.htm accessed November 2021.

LO 3.4

Use software tools to create summary statistics.

3.4 USING SOFTWARE TOOLS TO CREATE SUMMARY STATISTICS

In the previous section we introduced statistics that we will use throughout this book. At this point, it is helpful to learn how to calculate these statistics using Excel, Power BI, and Tableau. The chart in Exhibit 3.9 color-codes each calculation as green, orange, or red. The green items are straightforward in the tool of choice. The orange items require a bit of a workaround, but learning those workarounds is very useful. The red items are not easily done in the chosen software package.

	Excel Options: Function,/Formula, Data Analysis ToolPak (DA), and Pivot Table			Power BI	Tableau Options: Visualization and Summary Card (SC)	
	=Function() or =Formula	DA	Pivot Table		Visualization	SC
Mean	AVERAGE()	✓	Change Sum Values Default to Average.	Change Sum Default Value to Average.	Change SUM to AVG for Measure.	✓
Median	MEDIAN()	✓		Change Sum Default Value to Median.	Change SUM to MEDIAN for Measure.	✓
Mode	MODE.MULT()	✓	Count and Sort	Count and Sort	Count and Sort	✓
Skew	SKEW()	✓		DAX		✓
Range	MAX() – MIN()	✓	Create a measure using PowerPivot.	Create a Quick Measure.	Calculate max – min.	
IQR	QUARTILE.EXC(array,3)– QUARTILE.EXC(array,1) Or, create a box plot.			DAX or visualization library box plot	Subtract the values for Q1 from Q3 in a box plot.	
Standard Deviation	Sample: STDEV.S() Population: STDEV.P()	✓	Change Sum Values Default to StdDev or StdDevP.	Change Sum Default Value to Standard Deviation.	Change SUM to Std. Dev or Std. Dev (Pop.)	✓
Variance	Sample: VAR.S() Population: VAR.P()	✓	Change Sum Values Default to Var or VarP.	Change Sum Default Value to Variance.	Change SUM to Variance or Variance(Pop.)	

✓ = Included

DAX = Data Analysis Expressions

Array = a set of numbers, often used in the context of Excel functions to indicate that a calculation requires more than one single value.

Exhibit 3.9
Calculating Selected Summary Statistics with Excel, Power BI, and Tableau

Excel Functions

All Excel functions begin with an equal (=) sign, which is the signal to Excel to perform a calculation. Each set of parentheses should contain at least one argument. In Excel, an **argument** is a value that the function uses to perform calculations. With the exception of IQR (interquartile range), all of the Excel functions require that a range of cells be specified.

For IQR, a range of cells is specified followed by the quartile number. Exhibit 3.10 provides an example of using an Excel function to calculate the average sales price.

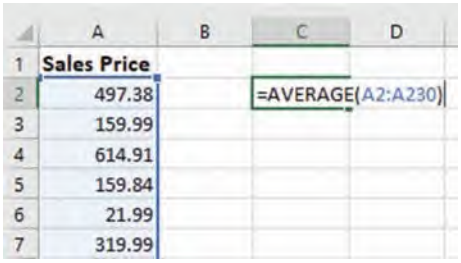


Exhibit 3.10
Excel: Calculating the Average Sales Price Using the AVERAGE() Function

Excel Data Analysis ToolPak—Descriptive Statistics

As noted in Chapter 2, Excel offers a Data Analysis ToolPak that helps you perform many of the business analytics explained in this text. If you aren't sure if you've installed the Excel Data Analysis ToolPak, look for it on the **Data > Analyze** ribbon (Exhibit 3.11). The appearance may vary by Mac, PC, and version of Excel. If you haven't installed the Data Analysis ToolPak yet, do so now following the instructions in Appendix E.



Exhibit 3.11
Finding the Data Analysis ToolPak in Excel

To generate descriptive statistics, select **Data > Analyze > Data Analysis**. A dialog box will open, as shown in Exhibit 3.12. Select Descriptive Statistics.

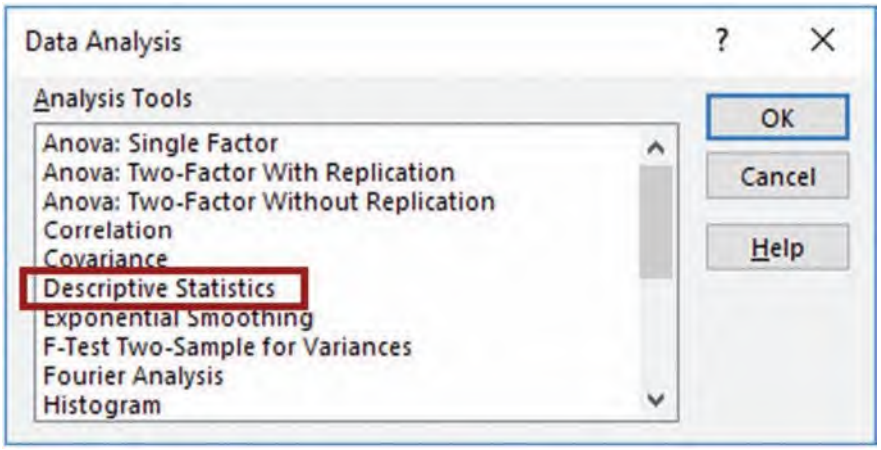


Exhibit 3.12
Selecting Descriptive Statistics from the Excel Data Analysis ToolPak

Descriptive Statistics generates basic statistics, including the mean (average), median, mode, skewness, standard deviation, minimum, and maximum of a data set as shown in Exhibit 3.13 for the salaries of a fictitious 10-person company, Tires-R-Us. The median and the mode are the same at $34,000, but the mean (at $40,600) is higher than the median and the mode because the maximum value of $100,000 skews the mean higher. The skewness is 3.16, indicating that the data are skewed to the right, or positively skewed. Recall that a perfectly symmetrical distribution has a skew value of 0. If the skew value is below 0, then the data set is skewed to the left, or negatively skewed. If the skew value is above 0, then the data set is skewed to the right, or positively skewed.

Exhibit 3.13
Descriptive Statistics from the Excel Data Analysis ToolPak

Salary	
Mean	$ 40,600.00
Standard Error	6600
Median	$ 34,000.00
Mode	$ 34,000.00
Standard Deviation	20871.03256
Sample Variance	435600000
Kurtosis	10
Skewness	3.16227766
Range	$ 66,000.00
Minimum	$ 34,000.00
Maximum	$100,000.00
Sum	$406,000.00
Count	10

Excel's kurtosis value is relative to the normal distribution; the closer that kurtosis is to zero, the closer the shape is to a normal distribution. We can also gain insight into our distribution based on whether kurtosis is negative or positive. A negative excess kurtosis suggests that the distribution is "flatter," so the values in the tails are less extreme. A positive excess kurtosis, like the one in Exhibit 3.13, suggests a "taller" or more peaked distribution, which implies that the values in the tails are more extreme. Kurtosis is particularly important when your analysis is related to risk management. Riskier distributions have positive kurtosis, whereas less risky distributions have negative kurtosis.

Changing the Default Aggregation in Excel Pivot Table, Power BI, and Tableau

Excel Pivot Tables, Power BI, and Tableau Visualizations all provide great tools to summarize data. They can aggregate data in a variety of ways, including Sum, Average, Count, Maximum, Minimum, and Standard Deviation.

Exhibits 3.14, 3.15, and 3.16 show how to change the default aggregate measure easily in an Excel Pivot Table, Power BI visualization, and Tableau visualization, respectively. In each example, the SUM aggregate for Sales Price is changed to AVERAGE, but you could also change the aggregate to a variety of other measures.

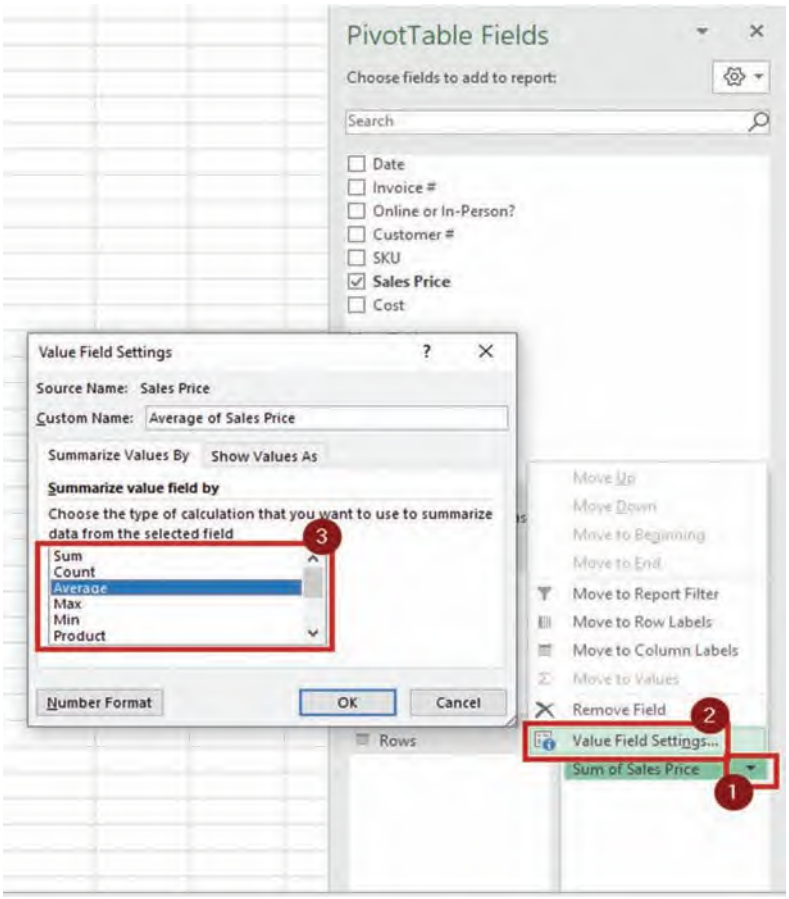


Exhibit 3.14
Changing the Default Value Field (Aggregation) Setting in an Excel Pivot Table

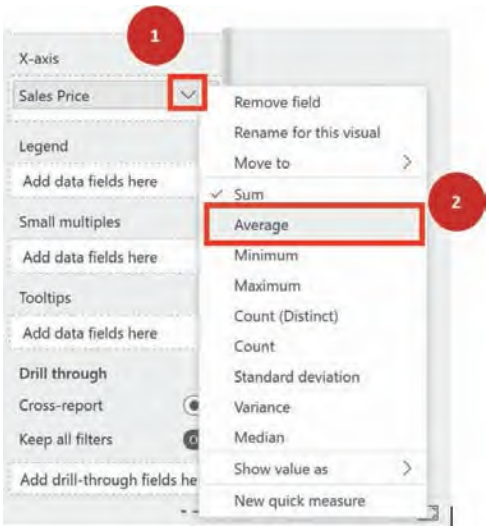


Exhibit 3.15
Changing the Default (Aggregate) Value in Power BI

Exhibit 3.16
Changing the Aggregate Default Measure in Tableau

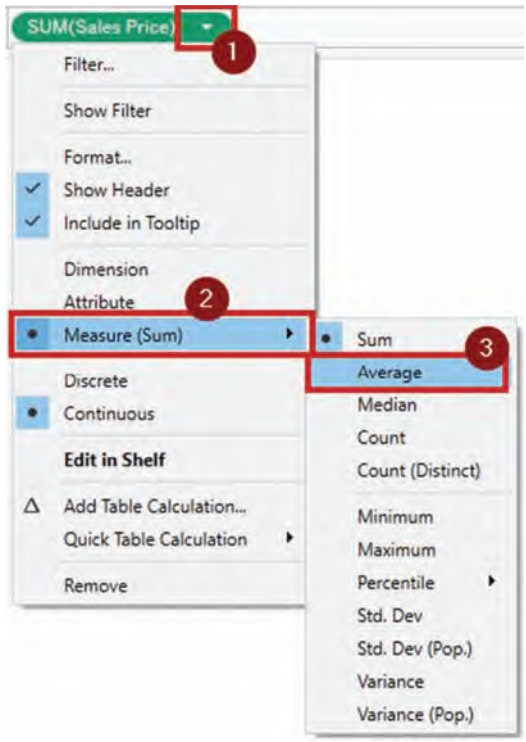
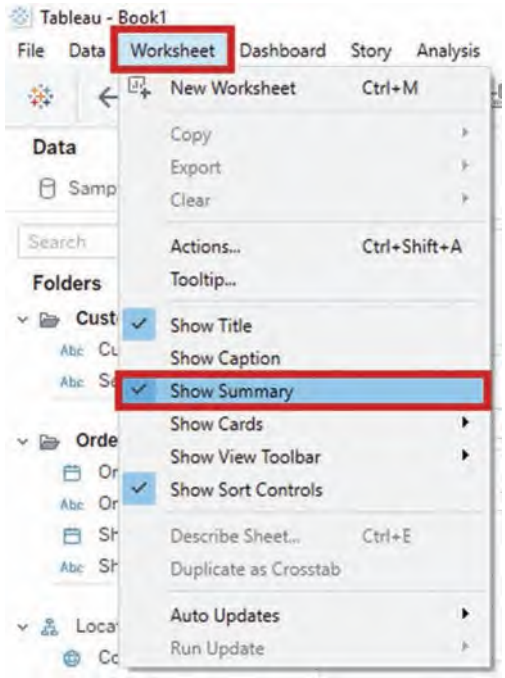


Tableau Summary Desktop Card

Like Excel's Data Analysis ToolPak Descriptive Statistics option, Tableau provides the ability to see a variety of summary statistics for any visualization you create, as shown in Exhibit 3.17.

Exhibit 3.17
Generating Summary Statistics in Tableau



PROGRESS CHECK

12. Of the tools listed in Exhibit 3.9, which tool (Excel Function/Formula, Excel Data Analysis ToolPak—Descriptive Statistics, Power BI, Tableau Visualizations, or Tableau Summary Card) can create all of the statistics mentioned in the exhibit?
13. Which two statistics can be easily created in all of the tools mentioned in Exhibit 3.9?

3.5 INTERPRETING AND VISUALIZING STATISTICS

LO 3.5

Interpret summary statistics and visualize them with histograms and box plots.

Whew—that's a lot of summary statistics! In later chapters and labs, you'll be interpreting and visualizing these statistics in the context of business decision-making. For now, though, let's put everything together to examine a complete set of summary statistics and what they tell us. Then we'll create two visualizations, histograms and box plots, which provide another way to view the statistics.

Exhibit 3.18 shows descriptive (summary) statistics output from Excel. The statistics are based on the sales price and cost data we used in Chapter 2 (see Exhibits 2.9 and 2.10). We have used boldfaced text and yellow highlights to emphasize the measures discussed below.

Sales Price		*Cost*	
Mean	194.3605	**Mean**	107.7226
Standard Error	12.24529	Standard Error	6.906444
Median	97.3	**Median**	52.58
Mode	59.99	**Mode**	354.66
Standard Deviation	185.3049	**Standard Deviation**	104.5135
Sample Variance	34337.91	Sample Variance	10923.06
Kurtosis	-0.10363	Kurtosis	0.185668
Skewness	1.147439	**Skewness**	1.226878
Range	592.92	**Range**	342.6
Minimum	21.99	**Minimum**	12.06
Maximum	614.91	**Maximum**	354.66
Sum	44508.56	Sum	24668.48
Count	229	Count	229

Exhibit 3.18 Example of Descriptive (Summary) Statistics (from Excel Data Analysis ToolPak)

It is important to look at each set of summary statistics individually, and also how they compare to each other. Let's begin by analyzing only the Sales Price data. Notice that the mean is higher than the median: This implies that some high outliers in this data set are skewing the data positively. Without digging into the underlying data, we won't know the precise cause, but we can assume that a few products are priced much higher than the majority of the products. The positive value for skewness supports that interpretation. We see the same pattern for mean, median, and skewness in the Cost summary statistics.

Now let's compare these two sets of descriptive statistics. The mean and median for Sales Price are higher than they are for Cost. This makes sense, because we should be selling our items for higher prices than their cost. We also see that the standard deviation and

the range are higher for Sales Price than Cost. This tells us that even though cost is a factor in determining sales price, it is clearly not the only factor. If it were the only factor, then the data sets would have similar levels of dispersion.

We can dig into the values that create the range by looking at the minimum and maximum values in each distribution. As expected, the minimum and maximum values are smaller for Cost than they are for Sales Price, and if we manually subtract each minimum from its respective maximum, we calculate the range. Because the measures of dispersion (variance and standard deviation) are higher for Sales Price than they are for Cost, we know that the data are more dispersed in the Sales Price data set than in the Cost data set.


DATA VISUALIZATION

Interpreting Statistics Through Visualizations

As we've seen throughout this text, transforming our data into a visualization can aid in interpretation and decision-making. We introduced some common visualization types in Chapter 1. Let's now examine two types of visualizations in more detail: histograms and box plots.

Recall from Chapter 2 that categorical data can easily be counted and grouped. Looking back at Exhibits 2.8 and 2.9, we can count the number of in-person sales versus online sales and easily group those observations. However, it isn't quite so obvious how to group numerical data. For example, how could we group Sales Price data? Or Cost data? We need to create "categories" in order to do so. Categories in numerical data are called *bins, classes,* or *intervals.* Both Excel and Tableau can auto-select bin sizes for you, but you can also override the auto-selection of the bin sizes and choose your own.

A **frequency distribution** is a table that uses bins or categories to list the frequency of various outcomes in a sample. It provides a quick snapshot of a data set's distribution. If we create nine bins for our data set, in a symmetric distribution we would expect the 5^{th} bin (the middle bin) to contain the most observations. We also expect the number of observations to decrease in a symmetric matter from the middle bin toward the lowest bin, and from the middle bin toward the highest bin. If the highest number of observations is in the 1^{st} or 2^{nd} bins (or in the 8^{th} or 9^{th} bins), that would mean the data are skewed.

Histogram

As we saw in Chapter 1, a **histogram** is a visual representation of a frequency distribution. It shows the frequency of the data using a display of rectangles whose areas are proportional to the underlying frequency of the data. The X-axis (or horizontal axis) in a histogram represents the bins, and the Y-axis (or vertical axis) represents the number of observations that fit in each bin. Exhibit 3.19 shows histograms for the Cost and Sales Price data we have been examining. Note that the two histograms have different scales, which are based on the size of the bin and the total range of the data. Each histogram also has reference lines for the average (or mean) and the median. These particular histograms were created in Tableau, but you can also use other tools to create histograms, including Excel.

Also note that a histogram looks similar to a column/vertical bar chart, but instead of the bars representing categories (such as product name, year, or percentage, as in Exhibit 1.9a in Chapter 1), the bars of a histogram represent bins. Also note that there are no gaps between the bars in a histogram, unless the gap represents the absence of data in that particular bin. In a bar chart, the Y-axis can be associated with a variety of descriptive statistics (average, sum, and count, for example), but the vertical axis in a histogram is always associated with the count of observations in each particular bin. For simplicity purposes, this textbook will use the term bar chart for both horizontal bar charts and vertical bar (or column) charts.

By examining the two histograms in Exhibit 3.19, we can perform an analysis similar to the analysis we performed when we looked at the measures numerically. We can see that

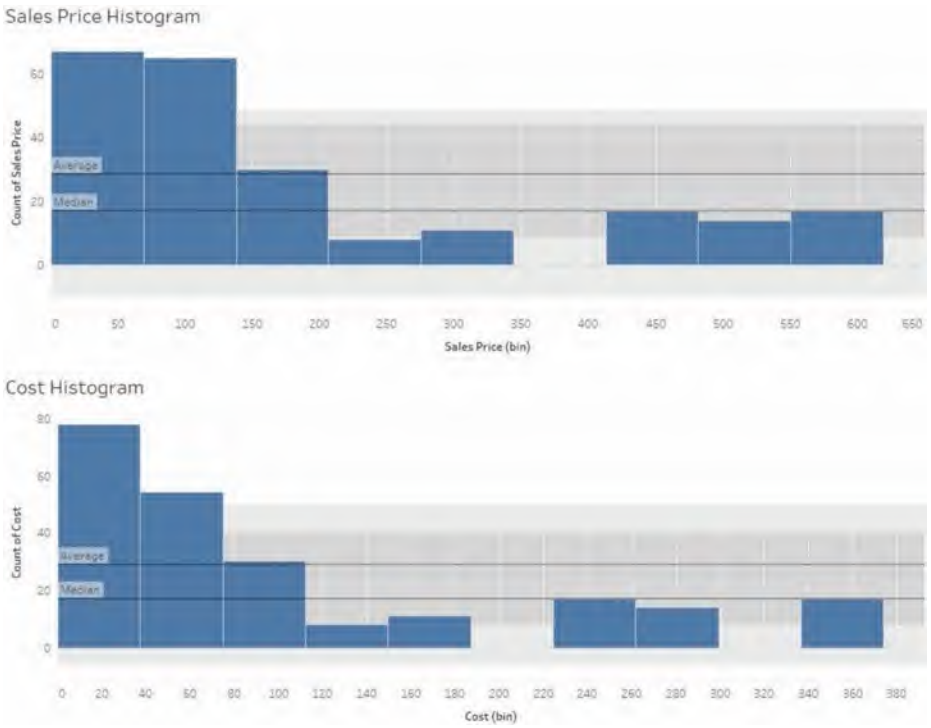


Exhibit 3.19
Using Histograms to Visualize Data

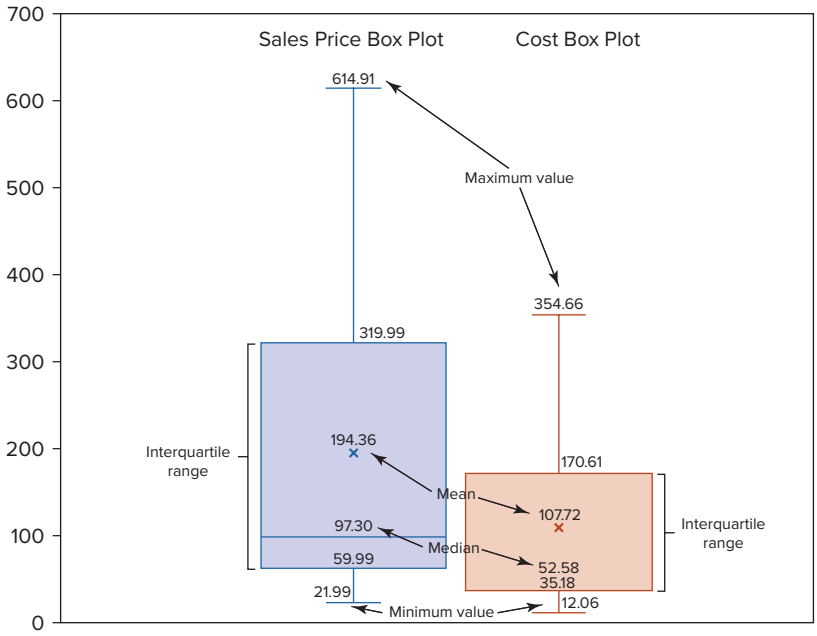
for each data set, the average or mean is higher than the median. The difference between those two measures indicates a skew in our data, but looking at the bins visually makes it much easier to immediately see the skew. There are far more items in the lower-value bins than in the higher-value bins. The next steps for this company certainly can't be discerned directly by comparing the two histograms; however, it is important to consider if there are follow-up questions based on these visualizations. For example, it may be worthwhile to look at the inventory records to see if the company is offering too many high-priced items that just don't sell.

Box Plots

We can use a **box plot** to visually represent how data are dispersed in terms of quartiles. Box plots are sometimes called **box-and-whisker plots** because they typically are made up of one box and two "whiskers." The box represents the interquartile range (Q1 to Q3), and the whiskers represent the minimum and maximum values within the lower/upper fences. The lower fence is defined as Q1 − 1.5IQR, while the upper fence is defined as Q3 + 1.5IQR. If your data have extreme outliers, they will be indicated as dots that extend beyond the whiskers. Note, there are no outliers in this example. Exhibit 3.20 shows two box plots: one for our Sales Price data and one for our Cost data. These box plots were created in Excel, and we added the labels to make it easier to read the different quartile values, means, and medians.

When we are comparing two different data sets on the same scale (in Exhibit 3.20, both Sales Price and Cost are on a 0-700 scale), a box plot shows us how the data sets compare. We can immediately see in Exhibit 3.20 some of the same observations we have made before—that the maximum and minimum are greater for Sales Price than for Cost, and that the mean is greater than the median for each data set. But by placing each box plot on the same scale, we can see the difference in ranges much more quickly.

Exhibit 3.20
Using Box Plots to Visualize Data



LAB CONNECTIONS

Labs 3.1 and 3.2 ask you to calculate summary statistics to compare online to in-person sales and visualize the data using a box plot. Lab 3.1 uses Excel, while Lab 3.2 uses your choice of Excel or Tableau.

PROGRESS CHECK

14. What does a histogram tell us about the data set?
15. How is a box plot different from a histogram?

LO 3.6

Explain confidence intervals and hypothesis testing.

3.6 HYPOTHESIS TESTING

The type of analysis you will perform depends on whether you already have a hypothesis in mind. As noted earlier in this chapter, a **hypothesis** is a proposed explanation (or supposition) made on the basis of limited evidence as a starting point for further investigation. If you do not yet know your data well enough to have formed a hypothesis, or if you plan to mine your data for trends or interesting observations, you will be performing *exploratory analysis,* which we explain in Chapter 4.

Often, though, you will approach your business analysis project with a question and a hypothesis in mind. In that case, you will be performing *confirmatory analysis* to confirm your hunch, as we explain in Chapter 5.

Hypothesis tests are used to determine if there are statistically significant differences between groups. Conducting a hypothesis test allows us to systematically compare the descriptive statistics of two groups or samples to determine if the difference is significantly greater than random variation. Hypothesis tests are also used in regression analysis to determine if a significant relationship exists between an input and outcome. This section focuses on hypothesis tests. We begin with a discussion of confidence intervals, which will help you understand hypothesis tests.

Confidence Intervals: Estimating Population Parameters from Sample Statistics

In explaining the differences between sample statistics and population parameters, we noted that it is usually easier to calculate sample statistics than to measure the entire population. We can certainly use a single sample statistic, such as the sample mean, to estimate what the mean of the entire population might be. The sample mean is an example of a **point estimate**, which is a single value (statistic) calculated from a sample that can be used to estimate a population parameter. Unless we are lucky, the mean will not equal the often unknown, true mean of the population.

Instead of using a point estimate to estimate a population parameter, we can use a confidence interval to estimate the population parameter. A **confidence interval** provides a range of numbers (an interval) around the point estimate constructed so that we are very confident that the interval contains the true population parameter at a certain level of confidence, such as 90% or 95%. That **confidence level** is determined by the analyst and reflects the probability that the true value of a population parameter falls within a certain range. A higher confidence interval (99%, for example) reflects a very high level of confidence that the true population parameter falls within the interval, but the range of that interval will be *much wider* than the range of a lower confidence interval of, say, 75%.

For example, consider living in a large city and planning to meet a group of friends for dinner. How certain are you that you will arrive at exactly 7:30 P.M. given the unknowns of weather, traffic, and other unidentified risks? Would you feel more confident telling your friends that you will probably arrive sometime between 7:15 and 7:45? That increased confidence is the essence of using a confidence interval instead of relying on one single value.

To create a confidence interval, we use a sample to calculate two numbers: the point estimate and the margin of error. Specifically, the confidence interval is equal to

$$\text{Point estimate} \pm \text{Margin of error}$$

The **margin of error** is the range of values above and below the sample point estimate, or the furthest that the sample point estimate might be from the population parameter. It is calculated as the desired confidence level plus the **standard error**. Error does not refer to a mistake, but instead refers to a gap between the population and sample results due to our inability to truly capture the population parameter. The standard error, which represents the dispersion in a sample distribution, is calculated using the standard deviation and the sample size. The smaller the standard deviation and the larger the sample size, the tighter the confidence interval will be, which should certainly encourage you to gather a large data sample.

Hypothesis Testing

Now that we understand confidence intervals, let's explain how hypothesis testing works. Let's suppose that managers hypothesize that the number of sales transactions at their stores

are higher on Saturdays than on Sundays. They want to know whether their hypothesis is correct because they need to decide whether to hire more staff to run the stores on Saturdays.

Step 1: Determine Hypotheses

Usually, a hypothesis test is composed of two hypotheses: the null hypothesis and the alternative hypothesis. The first is the base case or the statement being tested, often called the **null hypothesis**, abbreviated as $\mathbf{H_0}$. It assumes that the hypothesized relationship does not exist, and that there is no significant difference between two samples or populations. In our example, the null hypothesis is stated as:

Null hypothesis (H_0): Sales on Saturday are the same as sales on Sunday.

The **alternative hypothesis,** abbreviated as $\mathbf{H_A}$ is the statement being tested against the H_0. It states a potential result that the analyst may expect. When the null hypothesis is rejected, the alternative hypothesis is supported. In our example, the alternative hypothesis is stated as:

Alternative hypothesis (H_A): Sales on Saturday are not the same as sales on Sunday.

To test our hypotheses in this example, we use a **two-tailed hypothesis test** comparing Saturday sales to Sunday sales. We reject the null hypothesis if the difference between the two days is statistically significant. In other words, the alternative hypothesis is supported, indicating that Saturday sales are significantly greater than OR significantly lower than Sunday sales. Determining significance requires conducting a statistical test of the data comparing average Saturday sales to average Sunday sales.

If we are not interested in simply determining if there is a difference, but instead want to understand the particular *direction* of that difference (for example, greater than, less than, fewer, more), then we need to run a **one-tailed hypothesis test**. In this case, we adjust the hypotheses to the following:

Null hypothesis (H_0): Sales on Saturday are less than or equal to sales on Sunday.

Alternative hypothesis (H_A): Sales on Saturday are greater than sales on Sunday.

Exhibit 3.21
Example Hypotheses

Context	Null Hypothesis	Two-Tailed Alternative Hypothesis	One-Tailed Alternative Hypothesis
Weekend sales (Saturday compared to Sunday)	Sales on Saturday are the same as/less than or equal to sales on Sunday.	Sales on Saturday are not the same as sales on Sunday.	Sales on Saturday are greater than sales on Sunday.

Our hypotheses are summarized in Exhibit 3.21. Our next step is to determine whether to reject or fail to reject the null hypothesis. To make that decision, we must understand the concept of statistical significance.

Step 2: Set Statistical Significance (Alpha)

After we run our hypothesis tests, we decide to either *reject the null hypothesis* or *fail to reject the null hypothesis.* In our one-tailed Sunday sales example, if we reject the null hypothesis, our data support management's belief that sales on Saturday are greater than sales on Sunday. We can expect to see management hire more staff, such as cashiers and shelf stockers, to support Saturday sales. If we fail to reject the null hypothesis, then we accept that our data do not support management's belief and therefore management would not hire more Saturday staff. So, it is important to come to the correct conclusion.

How do we come to the correct conclusion? If we discover that the average sales from a sample of Saturdays are higher than the average sales from a sample of Sundays, simply finding that there is a difference between the two means is not enough. Instead, we need to determine if that difference is big enough not to have been due to chance, or to use statistical terminology, we need to determine if the difference is *statistically significant.*

The **significance level**, also called **alpha**, is the threshold that a business analyst determines is appropriate to measure the strength of the difference (or relationship) between two or more variables. When setting the significance level, we need to consider the correct decisions and the errors that can result from hypothesis testing. Depending on the significance level we choose, we may be at risk of the sample data misleading us into failing to reject the null hypothesis when we shouldn't—or to reject the null when we shouldn't. It's therefore important for us to consider the amount of risk we are most comfortable taking on in either direction—false positive or false negative.

A higher alpha or significance level (10%, for example) implies a higher probability of rejecting the null hypothesis when it is actually true in the population. In other words, with a higher alpha we are at a higher risk of of committing a Type I Error or false positive. When we are more comfortable failing to reject the null hypothesis when the alternative hypothesis is true in the population data, we are at risk of a **Type II error** or *false negative.* The higher the alpha level, the less likely a Type II error is to occur.

Exhibits 3.22 and Exhibit 3.23 summarize the four possible outcomes of a hypothesis test and how they might produce Type I and Type II errors and the resulting consequences.

Exhibit 3.22
Type I and Type II Errors

	Null Hypothesis Is Actually True in Population	Alternative Hypothesis Is Actually True in Population
Reject the Null Hypothesis	**Type I error (False positive)**	✓ Correct decision
Fail to Reject the Null Hypothesis	✓ Correct decision	**Type II error (False negative)**

Exhibit 3.23
The Consequences of Type I and Type II Errors

	Claim False: Sales on Saturday are less than or equal to sales on Sunday.	Claim True: Sales on Saturday are greater than sales on Sunday.
Reject Null: Sales on Saturday are greater than sales on Sunday.	Hire more ***unneeded*** staff on Saturdays.	Hire more staff on Saturdays.
Fail to Reject Null: Sales on Saturday are less than or equal to sales on Sunday.	Do not hire more staff on Saturdays.	Do not hire more ***needed*** staff on Saturdays.

Common alphas are 1%, 5%, and 10%. In the next section, we look more closely at what alpha means and how we use it to determine whether to reject or fail to reject the null hypothesis. We begin by examining an important statistical concept: the *p*-value.

Step 3: Calculate the Test Statistic

A test statistic is computed from data in a hypothesis test and is used to calculate a *p*-value, which determines whether the null hypothesis is rejected or not. There are many test statistics that can be calculated depending on the data and the hypotheses. In this section, we focus on a *t*-test.

A ***t*-test** is a statistical test used to determine if there is a significant difference between the means of two groups or two data sets. *T*-tests can compare independent groups or paired groups. For example, analysts might use a *t*-test for independent groups to compare the mean return on asset (ROA) for companies in the retail industry to the mean ROA for companies in the entertainment industry. A paired *t*-test (discussed in the next section) might compare the same group of companies at different times. For example, it might compare the mean ROA for companies in the retail industry in 2022 to the mean ROA of the same companies in the retail industry in 2023.

Let's continue examining our hypothesis that the number of sales transactions at stores is higher on Saturdays than on Sundays. We can run a *t*-test on data from the last year to determine if there is a difference in average number of sales transactions on Saturdays and Sundays. We can use different tools to run a *t*-test, such as Python, R, or Excel. A two-sample assuming unequal variances is the appropriate *t*-test to run because:

- For the sake of this example, we will assume that the company stores are only open twice a month on Sundays, but they are open every Saturday. Therefore, we have a different number of observations for Saturdays and Sundays, making a paired *t*-test not appropriate.
- The analysts are not sure about the variances of each sample and how they compare.
- Saturday sales data are independent from Sunday sales data. In other words, Saturday sales do not affect Sunday sales.

The output in Exhibit 3.24 shows the Excel output for this *t*-test. You can see that this output provides several measures, including a few summary statistics that we covered earlier in this chapter: the mean, variance, and count of observations for each group (Saturday and Sunday number of transactions).

Exhibit 3.24
Example of Output from a *t*-Test (from Excel Data Analysis ToolPak)

t-Test: Two-Sample Assuming Unequal Variances

	Saturday Sales	*Sunday Sales*
Mean	113.78	87.04166667
Variance	484.8281633	411.6068841
Observations	50	24
Hypothesized Mean Difference	0	
df	49	
t Stat	5.160451065	
P(T<=t) one-tail	2.22597E-06	
t Critical one-tail	1.676550893	
P(T<=t) two-tail	4.45193E-06	
t Critical two-tail	2.009575237	

The mean Saturday number of transactions from 50 observations is 113.78, and the mean Sunday number of transactions from 24 observations is 87.04. Clearly, these two means are different, but are they *significantly* different? *Significantly different* means that the difference is unlikely due to chance. We need two items to determine if the means are significantly different:

1. the test statistic, t Stat, which equals 5.16. Refer to Appendix D for the mathematical computation of the *t*-test (t Stat).
2. the *p*-value.

Step 4: Reject or Fail to Reject the Null Hypothesis Using the *p*-Value

We describe findings as statistically significant by interpreting the *p*-value of the test statistic. The ***p*-value** represents the probability that sample results occurred simply by chance as opposed to representing a pattern found in the population. The range for *p*-values is between 0 (no chance that the sample results are due to chance) and 1 (100% chance the results are due to chance). The smaller the *p*-value, the stronger the evidence is to reject the null hypothesis.

P-values are important, but we can do more to be accurate than simply looking for low *p*-values. Specifically, we can also compare the *p*-value to the alpha threshold. A result is statistically significant when the *p*-value is less than alpha, signifying that a difference between the two means was detected and that we can reject the default (null) hypothesis. The *p*-value results from a calculation that involves summary measures from your sample—it is completely dependent on the sample you are analyzing, and nothing else. Alpha is the only measure you control.

These are the guidelines for evaluating the results of your hypothesis tests:

If *p*-value > alpha: Fail to reject the null hypothesis. (Data do not present a significant result. That is, the means are not significantly different.)

If *p*-value ≤ alpha: Reject the null hypothesis. (Data present a significant result. That is, the means are significantly different.)

Returning to our example of Saturday versus Sunday sales, Exhibit 3.24 shows a one-tailed *p*-value of 2.23 E-06. This value is specified in *scientific notation.* Translate the letter "E" into "times 10 to the *x* power," where *x* is the number (positive or negative) that follows the E. The value of 2.23 E-06 is therefore 2.23×10^{-6}, which is a tiny number (= 0.00000222597), putting our *p*-value well below 0.01 percent. Even with an alpha (α) of 1%, this very small *p*-value is less than alpha, which means we reject the null hypothesis. We can state the results to management this way: "The test rejects the null hypothesis at a 1% level of significance, indicating that Saturday sales are significantly higher than Sunday sales." Graphically, the results look like Exhibit 3.25.

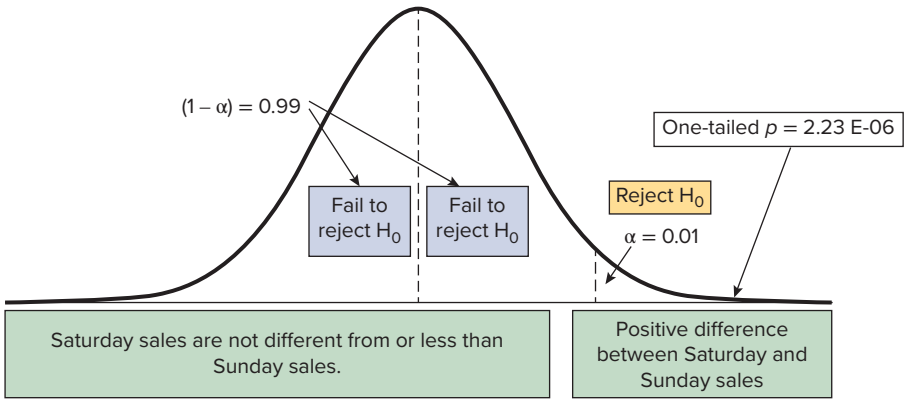


Exhibit 3.25 Evaluating the Results of a One-Tailed *t*-Test

Null hypothesis (H_0): Sales on Saturday are the same as sales on Sunday.

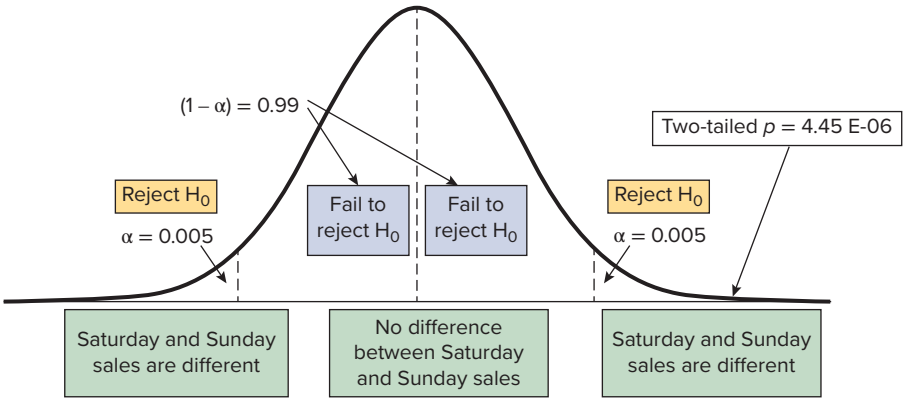
Alternative hypothesis (H_A): Sales on Saturday are less than or equal to sales on Sunday.

Alpha = 0.01.

Similarly, Exhibit 3.24 shows a two-tailed *p*-value of 4.45E-06, which is smaller than an alpha (α) value of 1%, which means we reject the null hypothesis. We can state the

results to management this way: "The test rejects the null hypothesis at a 1% level of significance, indicating that Saturday sales are not the same as Sunday sales." Given that the mean of Saturday sales is higher than the mean of Sunday sales, management can interpret the results as "Saturdays have more sales transactions than Sundays do." Graphically, the results look like Exhibit 3.26.

Exhibit 3.26
Evaluating the Results of a Two-Tailed *t*-Test



Null hypothesis (H_0): Sales on Saturday are the same as sales on Sunday.
Alternative hypothesis (H_A): Sales on Saturday are not the same as sales on Sunday.
Alpha = 0.01.

Management should then consider these results in deciding whether to invest in more staff to support Saturday sales. Certainly, there are more variables to consider, but based on this test alone, the data do support adding more staff on Saturdays because significantly more sales take place on Saturdays than on Sundays.

LAB CONNECTION

Lab 3.3 asks you to work in Excel with a *t*-test that compares the means of online sales and in-person sales.

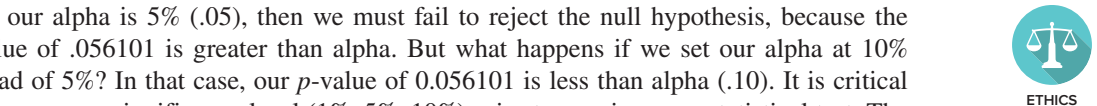
Now consider Exhibit 3.27, which shows the results of another hypothesis test. The *p*-value is highlighted in yellow.

Exhibit 3.27
Example of Two-Sample Hypothesis Test Assuming Unequal Variances

t-Test: Two-Sample Assuming Unequal Variances

	Group 1	*Group 2*
Mean	40.55345	33.64157
Variance	944.9389	610.8137
Observations	142	51
Hypothesized Mean Difference	0	
df	109	
t Stat	1.60131	
P(T<=t) one-tail	0.056101	
t Critical one-tail	1.658953	
P(T<=t) two-tail	0.112203	
t Critical two-tail	1.981967	

If our alpha is 5% (.05), then we must fail to reject the null hypothesis, because the *p*-value of .056101 is greater than alpha. But what happens if we set our alpha at 10% instead of 5%? In that case, our *p*-value of 0.056101 is less than alpha (.10). It is critical to choose your significance level (1%, 5%, 10%) prior to running your statistical test. The *p*-value shouldn't dictate which alpha you select, as tempting as that may be! In addition, it is important to rely on the results of your statistical analysis—even if the results do not align with what you hoped the outcome would be. The term *p-value hacking* refers to the practice of running analysis on various samples repeatedly until the desired results are obtained or analyzing several different variables and only publishing or sharing the results that were statistically significant. Hacking *p*-values is definitely not an ethical way to conduct hypothesis testing!



MINI CASE: Using Hypothesis Testing to Evaluate the Growth of an Investment In the Stock Market

Finance professionals are generally interested in achieving high returns on their investments. In search of these returns, they formulate hypotheses about what types of stocks will generate superior stock market returns. For example, does an annual investment of $1,000 have a better return if the investment is in small cap stocks? Mid cap? Large cap? Exhibit 3.28 shows the growth of an investment in stock market returns overall and by firm size for the years 1972 to 2018.

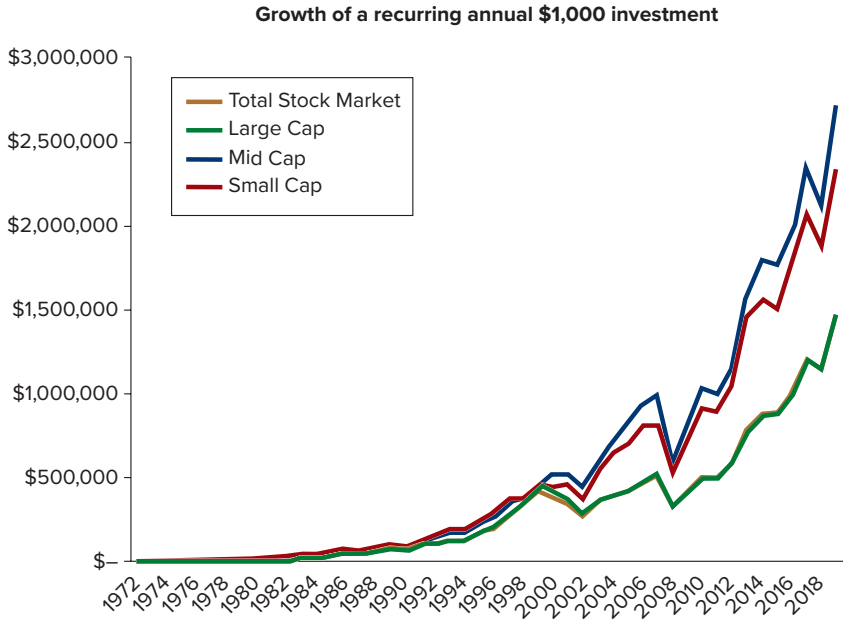


Exhibit 3.28
Investment in Stock Market Overall and by Size of Firm, 1972–2018
"Stock Returns: Small Cap vs. Mid Cap vs. Large Cap—Four Pillar Freedom," https://fourpillarfreedom.com/stock-returns-small-cap-vs-mid-cap-vs-large-cap. Accessed June 7, 2022.

Decision-Making with Data

If you wanted to test the hypothesis that large cap stocks have higher or lower returns than small cap stocks, what are the null and alternative hypotheses? What other hypotheses might finance professionals be interested in evaluating with hypothesis testing?

PROGRESS CHECK

16. Describe an occasion when an estimate using a confidence interval would be preferred over a point estimate.
17. Consider the statistics that factor into the calculation of the margin of error (confidence level, standard deviation, and sample size). Why do you think each of these items is important in having confidence about a population parameter falling within the calculated confidence interval?
18. Which hypothesis statement, the null or the alternative, is the base case? Which reflects the analyst's belief?
19. Describe the relationship between the selection of alpha (significance level) and the associated risk an analyst is willing to take on.
20. Is a Type I error more likely with an alpha (significance level) of 1% or an alpha (significance level) of 5%?
21. Which error, Type I or Type II, is a false positive?

LO 3.7

Understand which test statistic is appropriate for the data.

3.7 *t*-TESTS, ANOVA TESTS, AND CHI-SQUARE TESTS

Now that you have a basic understanding of hypothesis testing, let's look at some other test statistics that can be used.

t-Tests: Testing for a Difference of Means Between Two Paired (Dependent) Groups

In the prior section, the *t*-test treated each group as independent from the other. If the groups you are testing depend on each other, then you would run a different *t*-test specifically for paired groups. *Paired t-tests* are used to test for differences in:

- two samples where each observation in one sample can be paired with another observation in the other sample.
- before and after measurements for a sample (often called pre-test and post-test).

For example, suppose you are a manager concerned about your employees' understanding of workplace safety. If you give a pre-test to your employees to assess their understanding of workplace safety, then offer a training program on workplace safety and re-test them after the training program, you could run a paired *t*-test to determine if their understanding of workplace safety increased significantly after the training program. The two groups depend on each other because you want to keep each individual's scores on the pre-test and the post-test together so that you can monitor individual improvement.

In this scenario, our hypotheses are:

H_0: The post-test scores are lower than or equal to pre-test scores.
H_A: The post-test scores are higher than pre-test scores.

Exhibit 3.29 shows the first 10 records of data associated with a pre-test/post-test scenario. You can see that in this small sample, all individuals improved their performance on the test after participating in the training program. However, to be certain that the difference in performance is not simply due to chance, it is important to test the differences for significance using a *t*-test.

	pre-test	post-test
Participant 1	70	98
Participant 2	77	89
Participant 3	69	94
Participant 4	70	99
Participant 5	66	70
Participant 6	60	72
Participant 7	68	78
Participant 8	73	81
Participant 9	61	90
Participant 10	72	79

Exhibit 3.29
Ten Records from a Pre-Test/Post-Test Scenario

Exhibit 3.30 shows a sample output from a pre-test/post-test example of a paired *t*-test. You can interpret these results the same way you would interpret the results of an independent hypothesis test: Simply compare the appropriate *p*-value (one-tailed in this case) to the alpha you have selected. In Exhibit 3.30, note that the value for P(T<=t) one-tail is highlighted in red, and the value is a very small 2.79 E-16, which is lower than an alpha of 1%. We would therefore reject the null hypothesis. In other words, the training program significantly increased test scores.

***t*-Test: Paired Two Sample for Means**

	pre-test	*post-test*
Mean	70.18333	85.16667
Variance	40.45734	85.0904
Observations	60	60
Pearson Correlation	0.127731	
Hypothesized Mean Difference	0	
df	59	
t Stat	−11.0379	
P(T<=t) one-tail	2.79E-16	
t Critical one-tail	1.671093	
P(T<=t) two-tail	5.59E-16	
t Critical two-tail	2.000995	

Exhibit 3.30
***t*-Test Output for a Paired Sample**

ANOVA: Testing for Difference of Means Among Three or More Groups

As we've seen, a *t*-test compares the means between only two groups, whether they are independent groups or paired groups. When you want to compare the means across three or more groups, you can use the **analysis of variance (ANOVA) test.** The ANOVA output resembles the *t*-test output, and you determine statistical significance in the same way that you determine significance for the *t*-test: by comparing the calculated *p*-value to the alpha you've chosen. For example, you may use an ANOVA test when you want to compare average sales across three or more different markets or customer segments.

LAB CONNECTION

Lab 3.4 asks you to conduct an ANOVA test in Excel to determine if there is a difference in means and significant differences between the average four-year degree completion rate for public, private, and for-profit colleges.

Chi-Square Tests: Testing for a Relationship Between Two or More Variables

While the *t*-test and ANOVA test are used to determine differences between means, the **chi-square test** works with categorical data.[2] Specifically, it is used to test for the difference between observed and expected categorical data. The chi-square test is particularly helpful when you want to compare the distribution of a sample set of data to an expected distribution shape. As with hypothesis tests in general, you compare the calculated *p*-value to an alpha of your choosing.

Exhibit 3.31 summarizes the hypothesis tests and their purposes.

Test Name	Data Type	Number of groups	Purpose
Independent one-tailed *t*-test	Numerical	Two unrelated groups	To determine if a sample mean is significantly greater than or less than another sample mean
Independent two-tailed *t*-test	Numerical	Two unrelated groups	To determine if the means of two different samples are significantly different
Paired one-tailed *t*-test	Numerical	Two groups (or one group tested twice)	To determine if a sample mean is significantly greater than or less than a related sample mean
Paired two-tailed *t*-test	Numerical	Two groups (or one group tested twice)	To determine if the means of two related samples are significantly different
Analysis of variance (ANOVA) test	Numerical	Three or more unrelated groups	To determine if significant differences exist among three or more groups
Chi-square test	Categorical	Two or more groups	To compare the distributions of two different data sets and to compare observed versus expected data

Exhibit 3.31
Summary of Hypothesis Tests

MINI CASE: Evaluating the Returns to Quick Service Restaurant (QSR) Apps

As Exhibit 3.32 shows, customers are downloading apps for their favorite restaurants. By far the most popular app is **McDonald's**, with 2.2 million downloads in July 2021. Rounding out the top five are **Starbucks**, **Domino's Pizza**, **Chick-Fil-A**, and **Taco Bell**.

Why are restaurants interested in developing and maintaining their own apps? Do apps increase customer retention? Are customers who use apps more loyal than customers who do not? Do customers who use apps spend more than customers who do not?

[2]Chi is a Greek letter, and it is pronounced "kye."



Exhibit 3.32
Ten Most Downloaded Quick Service Restaurant (QSR) Apps in July 2021

Source: https://www.qsrmagazine.com/outside-insights/mcdonalds-has-most-downloaded-fast-food-app-america. Accessed November 2021.

Decision-Making with Data

How would you test the hypothesis that apps make more money for the restaurant? Would you use *t*-tests comparing sales revenue for customers who use the apps versus those who do not? How would you evaluate whether the company should continue to invest time, energy, and resources into such apps? What additional hypotheses would you generate, and how would you test them?

PROGRESS CHECK

22. A retail store operates in four different cities. As an analyst, you want to determine if there is a significant difference in the average sales during the past quarter across these four stores. Which hypothesis test should you use, and why?
23. What is the difference between a *t*-test and a chi-square test?

3.8 INTRODUCTION TO CORRELATION AND REGRESSION ANALYSIS

LO 3.8
Understand correlation and regression analysis.

As we examine our data, we are interested in addressing the question of how the variables are related. Is there a correlation—that is, do the variables move in coordination with each other? Or, does one variable affect the other variable? which allows us to predict future outcomes based on changes in the variables? To address these questions, we look at correlations and use regression analysis.

Correlation

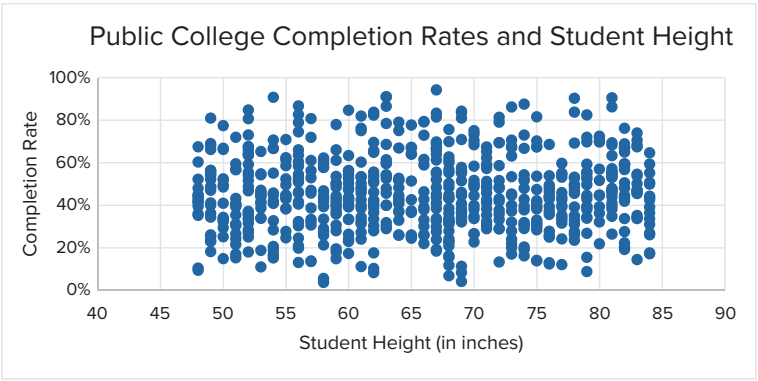
Correlation measures the linear association, or the relationship, between two variables by examining how the variables change with respect to each other. A **variable** is a measurement (or a measurable attribute) that changes over time or between individuals (or other subjects). Two variables' relationships may be categorized as having (1) no relationship, (2) a positive relationship, or (3) a negative relationship. As a starting point, a scatter plot of the data provides a visualization of which type of relationship exists.

For example, let's use the scatterplots in Exhibit 3.33 to look at public college complete rates in relation to average student height, course failure rate, and SAT scores. Exhibit 3.33(A) shows that there is no correlation (relationship) between the college completion rate and student height, which we would have expected.

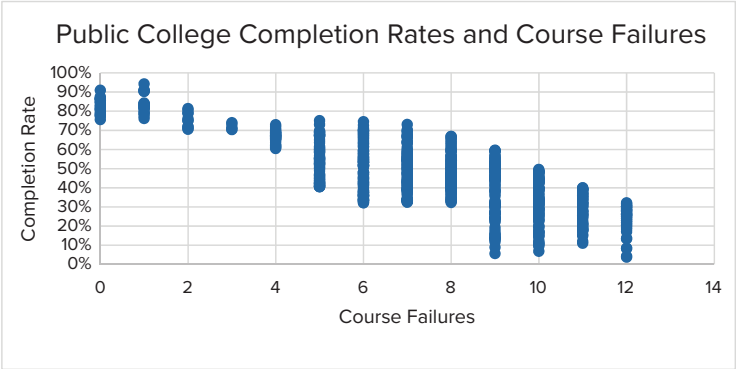
In contrast, the scatterplot of completion rates and course failures in Exhibit 3.33(B) indicates a negative relationship: The higher the course failures, the lower the completion rate. As we can see in Exhibit 3.33(C), the scatterplot for completion rates and high school SAT scores reveals a positive relationship: The higher the SAT scores, the higher the completion rate percentage.

From looking at the scatterplots, we cannot tell the degree to which the variables are related. But we can calculate the **correlation coefficient**, which mathematically defines the strength of the relationship. The correlation coefficient has a range of −1 to 1, with zero indicating no relationship. We can calculate correlation coefficients by using the Excel Data Analysis ToolPak and selecting Correlation, as shown in Exhibit 3.34.

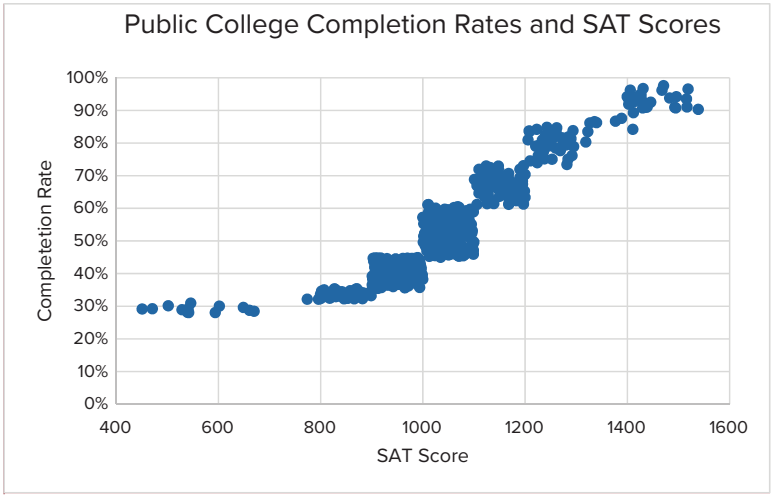
Exhibit 3.33
Scatterplots to Illustrate Correlation



(A) No Relationship



(B) Negative Relationship



(C) Positive Relationship

Exhibit 3.33 (Continued)

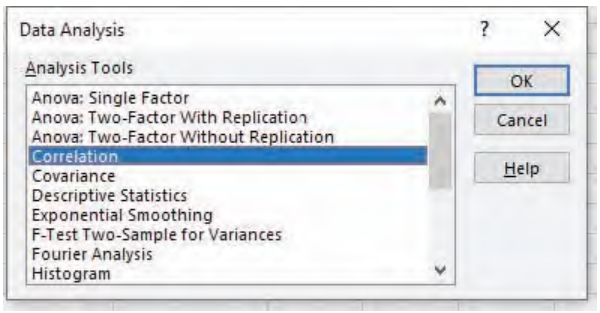


Exhibit 3.34 Selecting Correlation in the Excel Data Analysis ToolPak

The resulting computation of correlation coefficients is highlighted in yellow in Exhibit 3.35. The correlation coefficient of 0.04 for average height and completion rate is very close to zero and indicates that there is almost no measure of a relationship. The coefficient for completion rate and course failures is −0.80, indicating a strong relationship. The measures on the diagonal of the tables display the correlation of the variable with itself—a perfect correlation of 1.00.

	Average Height	*Completion Rate*
Average Height	1.00	
Completion Rate	0.04	1.00

	Course Failures	*Completion Rate*
Course Failures	1.00	
Completion Rate	−0.80	1.00

Exhibit 3.35 Correlation Coefficients Computed Using Excel Data Analysis ToolPak

Once we find a linear relationship between variables (using the correlation coefficient), our next step is to use regression. Regression determines whether variables move together, produces an equation of the line, and predicts expected outcomes based on the relationship. Simple regression allows us to examine a dependent (output) and one independent variable (input), while multiple regression allows us to examine a dependent variable and numerous independent variables.

Regression Analysis and Line of Best Fit

Sometimes managers want to know if a relationship exists between variables. To answer that question, business analysts can use **regression analysis**, which measures the relationship between one output variable and one or more input variables. If a relationship does exist, then we can use the regression model to help predict a given output variable. That output variable is called the **dependent variable**, as it is the variable that you are trying to predict based on one or more input variables. Those input variables are called the **independent variables**, they are the variables that you hypothesize relate to the dependent variable and can help explain or predict its outcome.

We can think about the results of regression analysis in terms of its specific function or model. In this chapter, we focus on linear regression, which is expressed as the following function:

$$y = mx + b$$

where

- y is the dependent variable—the item you are predicting,
- x is the independent variable(s)—the item(s) that help(s) to explain or predict y,
- m and b represent, respectively, the slope and the intercept of the line associated with the relationship between the dependent variable and the independent variables. The *intercept* is the point at which the line starts on the Y-axis. The *slope* is a number that describes the direction and steepness of the line. When running a regression, the slope is also called the *regression coefficient.*

Simple Regression

Let's look at a simple regression example (only one independent variable). Suppose we are considering the relationship between SAT scores and the college completion rate for first-time, full-time students at four-year institutions. We hypothesize that the higher the average SAT score at an institution, the greater the college completion rate will be for that institution. Here, SAT scores are the independent variable (SAT_AVG) and completion rate is the dependent variable. Exhibit 3.36 shows the output of the

Exhibit 3.36
Regression Results

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.80147835
R Square	0.64236754 (2)
Adjusted R Square	0.64208572
Standard Error	0.10486336
Observations	1271

ANOVA

	df	*SS*	*MS*	*F*	*Significance F* (1)
Regression	1	25.06431511	25.0643151	2279.33564	1.2641E-285
Residual	1269	13.95433624	0.01099632		
Total	1270	39.01865135			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.5741737 (4)	0.023762418	-24.163098	2E-106 (3)	-0.6207916	-0.5275557
SAT_AVG	0.00105984	2.21991E-05	47.7423882	1.264E-285	0.001016287	0.00110339

regression using Excel with the most significant parts of the output for our analysis highlighted and numbered.

1. The **Significance F** test evaluates whether the linear regression model provides a better fit of the data than a model that contains no independent variables. A very low significance F value (here, 1.26 E-285) indicates that the predictive model with SAT_AVG is better than an intercept-only model.
2. The **R Square** value measures goodness of fit, or how well the model predicted the dependent variable—in this case, completion rates. An R Square value of 0 means that the model has no ability to explain the dependent variable, while an R Square value of 1 means that the model perfectly explains the dependent variable. In this case, the R Square value is 0.642, meaning that *Holding all other variables constant, 64% of the variability in college completion time can be explained by the explanatory variable, SAT Average.*
3. The explanatory (or independent) variable of SAT_AVG has a very small *p*-value of 1.264 E-285. A very small *p*-value (that is less than your stated significance level) indicates that the independent variable is meaningful in predicting or explaining the dependent variable. A large *p*-value indicates that the independent variable is not meaningful in predicting or explaining the dependent variable. When there is only one explanatory variable, as in this case, the *p*-value for that explanatory variable will be identical to the Significance F value.
4. **Regression coefficients** are the quantities by which the variables in a regression equation are multiplied. In a simple regression, the regression coefficient is the slope of the line. Here the coefficient for SAT average equals a positive 0.00105984, indicating that the greater the SAT score is, the greater the college completion rate. The very small coefficient value indicates that for every one-point increase on the SAT test average, the average college completion rate will increase by approximately 0.001, or 0.1%.

In addition to interpreting the regression output, we can also view the data points and the line equation determined by the regression model: $y = 0.001x + (-0.574)$. The resulting line is sometimes called the **line of best fit**, because it best expresses the relationship between the actual data points. Exhibit 3.37 shows the actual data points in green and the line of best fit in red (starting the X-axis at 700).

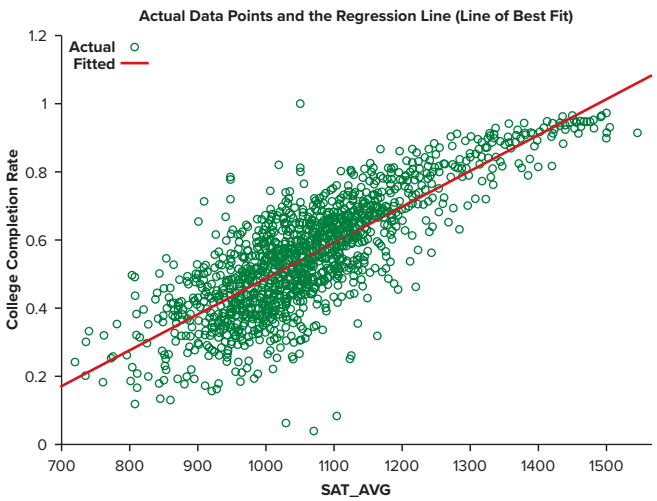


Exhibit 3.37
Regression Analysis: Line of Best Fit

Multiple Regression Analysis

The prior section showed an example of simple linear regression, which has only one independent variable. Perhaps you would be interested in determining if there are other variables, beyond SAT average, that influence the college completion rate. You could use multiple regression analysis, which works with more than one independent variable. Let's add Admission Rate (ADM_RATE) and the percentage of students on Pell Grants (PCTPELL) to our model. Exhibit 3.38 shows the results of the multiple regression analysis.

Exhibit 3.38
Multiple Regression Results

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.821018766
R Square	0.674071814
Adjusted R Square	0.673299473
Standard Error	0.100210174
Observations	1271

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	26.29308768	8.764362561	872.7637505	1.4932E-307
Residual	1266	12.71327206	0.010042079		
Total	1269	39.00635974			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.117994973	0.051702752	-2.282179728	0.022644028	-0.21942748	-0.016562467
ADM_RATE	-0.054161689	0.016801789	-3.223566729	0.00129834	-0.087124104	-0.021199274
SAT_AVG	0.000776711	3.46276E-05	22.4304279	4.28899E-94	0.000708778	0.000844645
PCTPELL	-0.327837866	0.029552687	-11.09333537	2.35075E-27	-0.385815496	-0.269860235

Multiple Regression Results

You can immediately see that the output is similar to that of the output for a regression analysis with only one explanatory variable. There are still a Significance F value and an R Square value, and there are coefficients and *p*-values for all three of the explanatory variables. Once again, the Significance F value is very small, and we can see that the R Square value increased slightly over the simple regression model, indicating that we improved the predictive power of the model. However, notice that in Exhibit 3.38 we have highlighted both the R Square value and the Adjusted R Square value.

When multiple explanatory variables are included in a regression analysis, the R Square value will automatically increase, even if the additional explanatory variables are not useful for the model. Thus, the **Adjusted R Square** is used for multiple regression models. The Adjusted R Square adjusts the R Square for each additional variable, increasing only if the additional variable improves the model.

Exhibit 3.39 summarizes the key statistics reported in the Excel output for regression analysis. It will be a useful reference when you encounter regression analysis in later chapters.

An important note: Neither correlations nor regression analysis provide a basis for determining cause and effect. Correlations measure the relationship between two variables and the degree to which they change with respect to each other. Regression analysis provides a line of best fit based on the data; this line can be used to predict outcomes given changes in the inputs.

Statistic	Purpose and Interpretation
Significance F	Answers the question: Does the linear regression model with independent variable(s) provide a better fit of the data than a model that contains no independent variables? **Interpretation:** A very low significance F value indicates that the predictive model including the independent variable(s) is significantly useful in predicting the dependent variable.
R Square	Used for simple (only one independent variable) regression. Answers the question: How much of the variation in the dependent variable is explained by the independent variable? **Interpretation:** An R square value of 0 means that the model has no ability to explain the dependent variable, and an R Square value of 1 means that the model perfectly explains the dependent variable.
Adjusted R Square	Used for multiple (more than one independent variable) regression. Answers the question: How much of the variation in the dependent variable is explained by the independent variables? **Interpretation:** The Adjusted R Square value increases only when an additional variable enhances the predictive power of the model.
Regression Coefficient	A separate regression coefficient is provided for each independent variable. Answers the question: For every one unit increase in the independent variable, how much will the dependent variable increase or decrease? (In multiple regression, interpret this response as "holding all other independent variables constant.") **Interpretation:** Use the regression coefficients to predict the numerical change in output based on a change in each input.
***p*-Value of Independent Variable**	In simple regression, this *p*-value is identical to Significance F. In multiple regression, the *p*-value is used to assess each individual explanatory variable for significance. **Interpretation:** Compare each *p*-value with your alpha to determine whether your data present a significant result.

Exhibit 3.39
Key Statistics in Regression Analysis: Purpose and Interpretation

LAB CONNECTION

Lab 3.5 asks you to perform a regression analysis in Excel. The goal is to derive cost drivers for activity-based costing, a common accounting practice.

PROGRESS CHECK

24. What is correlation?
25. Can you determine causation from correlations? From regression analysis?
26. What statistics can you use to determine if your regression model is able to explain the dependent variable?
27. In a regression, how do the Significance F and R Square statistics differ in terms of explaining the overall fit of the model?
28. What is the difference between R Square and Adjusted R Square in a regression model?
29. What does the *p*-value associated with the independent (or explanatory) variable tell us about the statistical importance of that independent variable in a regression?

CHAPTER SUPPLEMENT

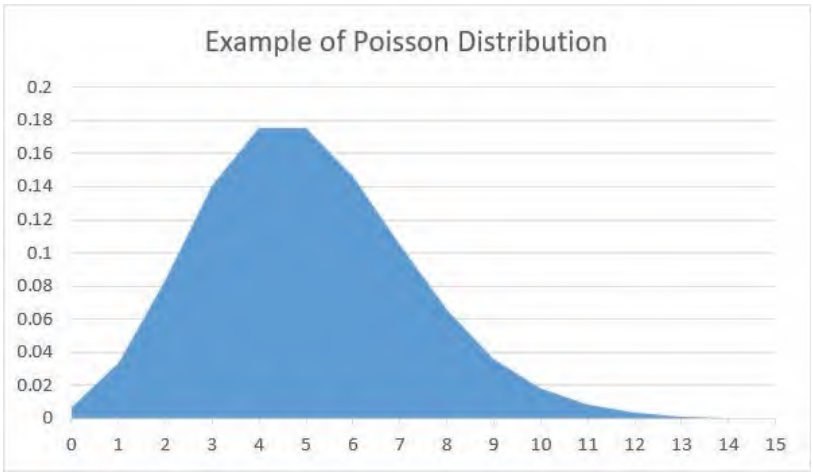
Discrete Data Distributions

We do not work with discrete data distributions in this textbook; we focus primarily on normal continuous distributions. However, it is useful to understand discrete data distributions to place the normal continuous distribution in context. Two examples of discrete data distributions are the binomial distribution and the Poisson distribution.

In scenarios that have a fixed number of trials with only two possible outcomes, the data will follow a **binomial distribution** pattern. Business examples of binomial distributions include determining whether customers who enter a store make a purchase, whether a website visitor signs up for a mailing list or not, and whether a stock price increased or not. A binomial distribution can be symmetric (like the normal and uniform distributions) or skewed, depending on whether there is an equal chance for each option to occur or not.

The **Poisson distribution** is used to model the number of occurrences per unit of something, such as number of defects per square feet or occurrences over the course of time. Other examples include the count of products sold during lunch hour or the count of defective products on an assembly line in a workday. Like the binomial distribution, the Poisson distribution works with discrete data, but it is applied to situations in which there are more than two possible outcomes. It has a low mean and is highly skewed to the right, as Exhibit 3.40 shows.

Exhibit 3.40
Example of a Poisson Distribution



Summary

In this chapter, we began our discussion of step 3 in the SOAR model: ANALYZE THE DATA. We looked at several helpful summary statistics, and we saw how those summary statistics can be visualized in histograms and box plots to help us understand the distribution of our data. Once you understand the distribution of your data, you can engage in more complex analysis. If you do not have any particular questions to answer or hypotheses to test, then you can perform exploratory analysis, which you will learn about in Chapter 4. If you have particular questions to answer, you will perform confirmatory analysis to test your

hypotheses, which you will learn about in Chapter 5. We also discussed core concepts in statistical testing, including significance levels, *p*-values, Type I and Type II errors, different types of hypothesis testing (*t*-tests, ANOVA tests, chi-square tests), and regression. You will encounter these concepts frequently throughout this book.

Key Terms

Adjusted R Square (*172*) A measure to describe how well a multiple regression (more than one independent variable) model predicts the dependent variable. Differs from the R Square measure by compensating for additional independent variables in the model.

alpha (*159*) The threshold that a researcher determines is appropriate to measure the strength of the difference (or relationship) between two or more variables.

alternative hypothesis (H_A) (*158*) The statement being tested against the null hypothesis ($\mathbf{H_0}$). It states a potential result that the analyst may expect.

analysis of variance (ANOVA) test (*165*) A statistical test used to analyze the differences in means among three or more groups.

argument (*148*) In Excel, a value that a function uses to perform calculations.

bias (*140*) Prejudice in favor of or against a thing, person, group, or idea.

binomial distribution (*174*) A distribution representing scenarios that have a fixed number of trials with only two possible outcomes.

box plot or **box-and-whisker plot (*155*)** A visual representation of data dispersion in terms of quartiles.

chi-square test (*166*) A statistical test used to test for the difference between observed and expected categorical data; particularly helpful when you want to compare the distribution of a sample set of data to an expected distribution shape.

cluster sampling (*138*) Random sampling only from specified groupings of the population.

confidence interval (*157*) A range of values estimated from sample data that contains the true population parameter at a certain level of confidence (typically 90% or 95%).

confidence level (*157*) The probability that the true population parameter falls within a certain range.

confirmation bias (*140*) Occurs when analysts analyze or present results in a manner that confirms their existing beliefs or theories while ignoring data and analyses that do not confirm their beliefs.

continuous data (*142*) Numerical data that can take on any numerical value, not just whole numbers, meaning there is an infinite set of values between any two observations in a continuous data scale, such as for weight or height.

convenience or **non-probability sampling (*138*)** Method of collecting data from convenient, easy-to-access data points.

correlation (*168*) A measure of the linear association, or the relationship, between two variables.

correlation coefficient (*168*) A measure of the strength of the relationship between two variables. The correlation coefficient has a range of −1 to 1, with zero indicating no relationship.

cost drivers (*211*) Any factor that causes a change in the overall cost of an activity.

data distribution (*141*) A function that specifies all of the possible values for a variable and how often they occur or could occur.

data reduction (*138*) The process of reducing the size of a data set to something more manageable and suitable for a business analysis project.

dependent variable (*170*) In regression analysis, the outcome that you are trying to predict based on input (independent) variables.

descriptive or summary statistics (*137*) Measures that characterize, summarize, and organize features and properties of the data to facilitate an understanding of business results and the underlying data.

discrete data (*142*) Numerical data that only take on whole-number (integer) values and have a finite set of values between any two observations.

filtering (*138*) The process of choosing a smaller portion of a data set for subsequent viewing or analysis.

frequency distribution (*154*) A table that uses bins or categories to list the frequency of various outcomes in a sample.

histogram (*154*) A visual representation of a frequency distribution; it shows the frequency of the data using a display of rectangles whose areas are proportional to the underlying frequency of the data.

hypothesis (*137*) A proposed explanation (or supposition) made on the basis of limited evidence as a starting point for further investigation.

hypothesis tests (*157*) Statistical tests of an assumption or theory, specifically to determine if there are significant differences between groups.

independent variable (*170*) In regression analysis, the input variables that you expect to be useful in predicting or explaining a dependent variable.

inference (*136*) A conclusion drawn on the basis of evidence and reasoning about the characteristics of a population based on sample characteristics.

inferential statistics (*137*) Measures that are calculated using a sample of the desired population. These measures are used to test hypotheses regarding the population.

interquartile range (*145*) The midsection of a data distribution, or the range between second and third quartiles.

kurtosis (*144*) A measure used to describe a data distribution; it describes the thickness of the tails of a distribution, specifying whether values are more clustered in the peak of a distribution or in the tails. (The tails are the ends of the distribution on either side.)

line of best fit (*171*) A line in the format $y = mx + b$ that best expresses the relationship between the actual data points. This line is determined by a regression model.

margin of error (*157*) The range of value above and below the sample point estimate, or the furthest that the sample point estimate might be from the population parameter. Calculated as confidence level plus standard error.

mean (*142*) The average of a set of measurements in a data set.

measures of central tendency (*142*) Measures that describe the center point of a data set, such as the mean (average), median, or mode.

measures of dispersion (*144*) Measures that describe how varied, or dispersed, a data set is.

median (*142*) The midpoint of a distribution.

mode (*142*) The most common observation in a data set.

nonresponse bias (*140*) The partiality (or bias) that results when respondents differ from non-respondents.

non-probability sampling (*138*) Method of collecting data from convenient, easy-to-access data points.

normal (Gaussian) distribution (*146*) A bell-shaped probability distribution that is symmetric about its mean, with the data points closer to its mean more frequent than the data points further from its mean.

null hypothesis (H_0) (*158*) In hypothesis testing, the base case with the assumption that the hypothesized relationship does not exist, and that there is no significant difference between two samples or populations.

outlier bias (*140*) The potential disproportionate influence of outliers in study results.

one-tailed hypothesis test (*158*) A hypothesis test to determine the direction of the difference (greater than or less than) between two groups.

parameter (*136*) A characteristic of a population.

percentile (*145*) A value on a scale of 100 that indicates the percent of a distribution that is equal to or below it.

point estimate (*157*) A single value (statistic) calculated from a sample that can be used to estimate a population parameter.

Poisson distribution (*174*) A distribution that models count data over the course of time. It is similar to the binomial distribution in that it describes discrete data, but it is applied to situations in which there are more than two possible outcomes.

population (*136*) A group of phenomena that have something in common.

probability distribution (*141*) A statistical function that describes the possible values in a population and the likelihood that any given observation (random variable) can take a given range or value.

***p*-value (*161*)** The probability that sample results occurred simply by chance as opposed to representing a pattern found in the population.

random variable (*141*) A variable used to quantify outcomes of random occurrences.

range (*144*) A measure of dispersion; the difference between the maximum value and the minimum value in a data set.

regression analysis (*170*) A statistical analysis used to predict if a relationship exists between variables; measures the relationship between one output variable and various input variables.

regression coefficient (*171*) The amount by which the dependent variable will increase or decrease (depending on the sign of the coefficient) with a 1-unit increase in the independent variable. When there are multiple independent variables, the regression coefficient is interpreted per variable, with the assumption that all other variables are held constant. In a simple regression, the regression coefficient is the slope of the line.

R Square (*171*) A measure to describe how well a simple regression (one independent variable) model predicts the dependent variable.

sample (*136*) A subset of members of a population; researchers often select samples that they believe represent the population.

selection bias (*140*) The bias introduced by the selection of individuals or data that are likely to sway the results of a study in favor of the analyst's beliefs or hypothesis.

Significance F (*171*) A test that evaluates whether a linear regression model provides a better fit of the data than a model that contains no independent variables.

significance level or **alpha (*159*)** The threshold that a researcher determines is appropriate to measure the strength of the difference (or relationship) between two or more variables.

simple random sampling (*137*) A sample in which every observation in the population has an equal chance of being selected into the sample.

skewness (*143*) The extent to which data are asymmetric about their mean.

statistic (*136*) A characteristic of a sample.

standard deviation (*145*) The square root of the variance; the dispersion of a data set relative to its mean.

standard error (*157*) A calculation that represents the dispersion in a sample distribution. Used for calculating margin of error, which is calculated as confidence level + standard error.

standard normal distribution (*146*) A data distribution that has a mean, median, and mode of 0 and a standard deviation of 1.

stratified random sampling (*138*) Random sampling from known subgroups to ensure adequate representation in the sample.

summary statistics (*137*) Measures that characterize, summarize, and organize features and properties of the data to facilitate an understanding of business results and the underlying data. The basic descriptive statistics include the mean, standard deviation, minimum, and maximum of a data set.

symmetrical (*143*) Used to describe a data set in which the mean, median, and mode are equal. Symmetrical data distributions have an equal number of values on either side of the middle point of a distribution.

Type I error (*159*) A false positive that occurs when the analyst incorrectly rejects a true null hypothesis.

Type II error (*159*) A false negative that occurs when the analyst fails to reject a null hypothesis that is false.

***t*-test (*160*)** A statistical test used to determine if there is a significant difference between the means of two groups or two data sets.

two-tailed hypothesis test (*158*) A hypothesis test to determine if there is a significant difference between two groups or two data sets, regardless of the direction of the difference (greater than or less than).

uniform distribution (*147*) A probability distribution in which all outcomes are equally likely.

variable (*168*) A measurement (or a measurable attribute) that changes over time or between individuals (or other subjects).

variance (*145*) The average of the squared deviations from the mean for each observation; a measure of dispersion.

***z*-score (*146*)** A statistical indicator of how many standard deviations a data point (or observation) is from its population mean.

ANSWERS TO PROGRESS CHECKS

1. A population includes every item of interest in a group of phenomena that have something in common. A sample is a subset of that population. It is preferable (but generally not feasible) to survey an entire population because if every item is included in the study, we can be more certain that our results are correct. The use of a sample allows us to draw inferences with some uncertainty.
2. The impact will be considerable. There will be more faith in the results of the audit because the results will be based on the entire data set of transactions rather than on just a subset.
3. A parameter is a characteristic of a population, while a statistic is a characteristic of a sample. The numerical values may differ if the sample is not fully representative of the population, regardless of how well we attempted to make our sample reflect the population as a whole.
4. A hypothesis is a proposed explanation and is a starting point for further investigation.
5. All three sampling methods are used to select a subset of a population. In a simple random sample, every member of the population has an equal opportunity to be selected. In stratified and cluster sampling, not every member of the population has an equal opportunity to be selected into the sample. These two methods require you to split the population into groups and select your sample based on the proportions of each group. In stratified sampling, representatives from every group will be included; in cluster sampling, some groups will be systematically left out.
6. In convenience sampling, measures are not taken to ensure the sample is representative of the population being studied. A convenience sample could be simply the first 10 people who enter a store on a Monday morning or the first 500 people who take a survey on a public website. Even though convenience sampling is not likely to result in a representative sample, it is sometimes used when time is tight and budgets are low.
7. Nonresponse bias can occur when people are given a choice whether to respond or not. A researcher can eliminate some of that bias by providing some type of incentive for participating in a survey or the experiment (for example, a chance to win a gift card). Nonresponse bias can also result from poorly designed surveys or survey questions that respondents consider intrusive or uncomfortable. You can reduce nonresponse bias by not including such questions on your survey.
8. Selection bias occurs when the researcher purposefully selects portions of the population that are likely to sway the results of a study in favor of the researcher's beliefs or hypothesis. You can avoid selection bias by taking care not to select only those participants who are likely to answer in one particular way over another. Simple random sampling and stratified sampling are the best way to avoid selection bias.
9. Understanding the different measures of central tendency (mean, median, mode) can help you understand if your data are skewed or not, which is helpful in recognizing if the data set contains extreme high or low values. Using range or standard deviation to understand the dispersion of your data helps you determine whether most of the data points are clustered around the mean or if they are more spread out, which helps you recognize if values farther from the mean are truly extreme or not.
10. A value of 0 for skewness represents a symmetrical data set, and the mean, median, and mode should all be identical. A negative value for skewness indicates that the data are skewed to the left, or negatively skewed, and a positive value for skewness indicates that the data are skewed to the right, or positively skewed.
11. The standard deviation is a measure of dispersion. If Data Set A has a higher standard deviation than Data Set B, then the data in Data Set A are more varied (spread out farther away from the mean) than the data in Data Set B.
12. Excel Function/Formula
13. Mean and standard deviation

14. A histogram helps us identify how dispersed the data set is and if there is a skew.
15. A box plot is different from a histogram in several ways. A histogram is a visual representation of a frequency distribution; it shows the frequency of the data using a display of rectangles whose areas are proportional to the underlying frequency of the data. A box plot provides more descriptive statistics. It is based around quartiles, and particularly the interquartile range. It also clearly denotes outliers. A box plot is particularly useful for comparing two or more distributions.
16. When you are attempting to estimate a population parameter, a confidence interval will nearly always be preferred over a point estimate simply because it provides room for error in the sample statistic. The sample statistic is based only on the sample of data you gathered, and it is rather risky to assume that the exact mean of a sample is identical to the mean for the population as a whole. A confidence interval is more likely to contain the true parameter.
17. The confidence level is chosen by the analyst—it is important because it specifies how confident the analyst is that the resulting interval contains the population parameter. An analyst will not always want to choose the highest level of confidence, though, because there is a tradeoff: The higher the confidence level, the broader the interval. The standard deviation is important to include because the more varied the data (thus the higher the standard deviation value), the broader the interval will be. The sample size has the opposite effect: The larger the sample size, the narrower the interval will be.
18. The null hypothesis is the base case. The analyst's belief is the alternative hypothesis.
19. Common alpha (or significance levels) are 1%, 5%, and 10%. The higher the alpha level, the less likely a Type II error is to occur. Conversely, the lower the alpha level, the less likely a Type I error is to occur. Analysts need to consider which risk they are more comfortable taking—a higher potential for a false negative (Type II error) or a false positive (Type I error).
20. A Type I error is more likely with an alpha of 5%.
21. A false positive is a Type I error.
22. ANOVA test, because you are comparing three or more data sets.
23. A chi-square test is used with categorical data to compare the distributions of two different data sets and to compare observed versus expected data. A *t*-test is used with numerical data to determine if the means of two different samples are significantly different.
24. Correlation measures the linear association, or the relationship, between two variables by examining how the variables change with respect to each other.
25. Neither correlations nor regression analysis provide a basis for determining cause and effect. Correlations measure the relationship between two variables and the degree to which they change with respect to each other. Regression analysis provides a line of best fit based on the data; this line can be used to predict outcomes given changes in the inputs.
26. The Significance F or the R Squared value.
27. The Significance F and R Square statistics represent similar concepts, with both describing the predictive power of the model. Significance F is interpreted in the same way that you interpret *p*-values for hypothesis tests, and it indicates whether the relationship between the independent variables and the dependent variable is due only to chance (which would result in a relatively higher Significance F value) or if there is likely a significant relationship between the variables (resulting in a Significance F value lower than alpha). A very low significance F value indicates that the predictive model created by the independent variables is indeed significantly useful in predicting the dependent variable. The R Square value measures how well the model predicts the dependent variable, and it can be any value between 0 and 1. An R Square value of 0 means that the model has no ability to explain the dependent variable, while an R Square value of 1 means that the model perfectly explains the dependent variable.

28. Both R Square terms evaluate the effects of one or more independent variables on a dependent variable (that is, the percentage of variation explained by the independent variables). The Adjusted R Square takes into consideration whether the additional input variables are contributing to the model. In a simple regression, the two measures are the same.

29. The *p*-values for independent variables are the probability of explaining the dependent (outcome) variable from the independent variable alone. The *p*-value helps to explain how useful a given independent variable is in a model. If your regression model has only one independent variable (a simple regression model), then the *p*-value will match the Significance F. If there are multiple independent variables in your regression model (a multiple regression model), then the *p*-values will be unique for each separate independent variable.



Multiple-Choice Questions

1. (LO 3.1) Which of the following is a characteristic of the population?
   a. Parameter
   b. Statistic
   c. Type I error
   d. Regression
2. (LO 3.5) Which of the following show how data sets compare to each other with a visual representation of data dispersion by quartiles?
   a. Bar graphs
   b. Box plots
   c. Histograms
   d. Line charts
3. (LO 3.6) If *p*-value $\leq$ alpha threshold, you should
   a. reject the null hypothesis; the result is significant.
   b. fail to reject the null hypothesis; the result is not significant.
   c. reject the null hypothesis; the result is not significant.
   d. fail to reject the null hypothesis; the result is significant.
4. (LO 3.3) Which quartile is the cutoff for the median?
   a. Quartile 1
   b. Quartile 2
   c. Quartile 3
   d. Quartile 4
5. (LO 3.7) Which of the following is used to test for a difference of means between three or more groups?
   a. *T*-test
   b. Regression analysis
   c. ANOVA test
   d. Correlation

6. (LO 3.7) Which statistical test is used to determine how closely observed data fit the expected data?
   a. Chi-square
   b. Regression
   c. ANOVA
   d. Correlation
7. (LO 3.8) Which of the following is used to predict a specific dependent variable value based on independent variable inputs?
   a. Correlation
   b. Regression analysis
   c. *T*-test
   d. ANOVA test
8. (LO 3.6) Which of the following is the base case with the assumption that the hypothesized relationship does not exist, that there is no significant difference between two samples or populations?
   a. Null hypothesis
   b. Alternative hypothesis
   c. Original scenario
   d. Default assumption
9. (LO 3.2) Every observation in the population has an equal chance of being selected into the sample in which type of sampling method?
   a. Cluster sampling
   b. Stratified sampling
   c. Simple random sampling
   d. Convenience sampling
10. (LO 3.4) What is the name of the tool that you can use for quickly calculating statistics in Excel (other than Excel functions)?
    a. Summary Card
    b. Descriptive Statistics
    c. Regression
    d. *T*-test

Discussion Questions

1. (LO 3.3) What is a limitation of range as a measure of dispersion?
2. (LO 3.3) Why are standard deviation and the interquartile range better measures of dispersion than range?
3. (LO 3.5) When comparing data sets, which visualization is more helpful, a histogram or a box plot? Why?
4. (LO 3.6) If the *p*-value of the test statistic in a hypothesis test equals 0.04 and alpha is set at 5%, do you reject or fail to reject the null hypothesis?
5. (LO 3.6, 3.7) What is the difference between an ANOVA test and a *t*-test?
6. (LO 3.5) What is the difference between a bar chart and a histogram? If you want to create a visualization of the total number of employees for each company department, would you use a bar chart or a histogram? Explain.

7. (LO 3.5) Managers want a visualization showing the total number of employees aggregated by the following age groups: < 18, 18–30, 31–50, 51–65, > 65. To create this visualization, would you use a bar chart or a histogram? Why?
8. (LO 3.1) To prepare its payroll, a company requires all of its employees to submit a weekly, digital time sheet showing hours worked broken down by task. Are the data collected by the time sheets an example of a population or a sample? Would the data analysis calculate and use parameters or statistics?
9. (LO 3.2) Assume that a company wants to collect customer survey data by state. The company knows that all customers will not respond. It also wants to make sure that responses are included from customers in all states. What method of sampling discussed in this chapter would be the best for the company's purpose? Why?
10. (LO 3.4) How would you simultaneously generate the mean, median, mode, standard deviation, and skewness for a set of data in Excel?
11. (LO 3.8) If you wanted to predict expected sales revenue from advertising expenditures, would you use simple or multiple regression analysis? What would be the independent variable, and what would be the dependent variable?



Brief Exercises

1. (LO 3.1, 3.3, 3.5) Match each statistical term to its definition.

Statistical Term	Description
1. interquartile range	a. A characteristic of a population
2. histogram	b. A characteristic of a sample
3. parameter	c. A visual representation of a frequency distribution
4. statistic	d. The difference between the maximum value and the minimum value in a data set
5. range	e. The midsection of a data distribution

2. (LO 3.1, 3.4, 3.5, 3.6, 3.8) Match each business analytics statistical function in Excel's Analysis ToolPak to its use.

Business Analytics Statistical Function	Use
1. correlation	a. To determine if there is a significant difference between the means of two groups or two data sets
2. regression	b. To understand if and the extent to which variables are related to each other
3. descriptive statistics	c. To understand the basic statistics, including the mean, standard deviation, minimum, and maximum of a data set
4. histogram	d. To understand the frequency of the data using a display of rectangles with area proportional to the underlying frequency of the data
5. *t*-tests	e. To understand the relation between specific dependent variable values and independent variable inputs

3. (LO 3.6, 3.7, 3.8) Match each hypothesis test to its description.

Hypothesis Test	Description
1. ANOVA	a. A statistical test used to determine how closely observed, or realized, data fit the expected data
2. chi-square test	b. Statistical analysis used to understand the relationship between a dependent variable output and independent variable inputs
3. *t*-test	c. A statistical test used to determine if there is a significant difference between the means of two groups or two data sets
4. regression	d. A statistical test used to analyze the differences in means among three or more groups

4. (LO 3.3) Match each term to its definition.

Term	Definition
1. median	a. The extent to which data are asymmetric about their mean
2. symmetrical	b. Describes a data set in which the mean, median, and mode are equal
3. range	c. The average of a distribution
4. standard deviation	d. The difference between the maximum value and the minimum value in a data set
5. mean	e. The dispersion of a data set relative to its mean
6. skewness	f. The midsection of a data distribution, or the range between the 25^{th} percentile and the 75^{th} percentile
7. interquartile range	g. The midpoint of a distribution

5. (LO 3.2) Match each term with its definition.

Sampling Techniques and Biases	Definitions
1. simple random sampling	a. A sampling method in which every observation in the population has an equal chance of being selected
2. stratified random sampling	b. Method of collecting data based from convenient, easy-to-access data points
3. non-probability (convenience) sampling	c. Random sampling from known subgroups to ensure adequate representation
4. cluster sample	d. Random sampling only from specified groupings of the population
5. confirmation bias	e. The partiality (or bias) that results when respondents differ from nonrespondents
6. outlier bias	f. The potential disproportionate influence of outliers in study results
7. nonresponse bias	g. Occurs when the analyst analyzes or presents new evidence as confirmation of existing beliefs or theories

Problems

1. (LO 3.1, 3.6) Answer the following questions about hypothesis testing. Assume that:

   H_o: The company's ROA is the same as the industry's average ROA.

   H_A: The company's ROA is not the same as the industry's average ROA.

   a. If the alpha threshold is 0.05, and the *p*-value on a *t*-test comparing a company's ROA to the industry average ROA is 0.045, what do you conclude about the null hypothesis? Is the ROA the same as or different from that of the industry?

   b. If the alpha threshold is 0.10, and the *p*-value on a *t*-test comparing one company's ROA to the industry average ROA is 0.11, what do you conclude about the null hypothesis? Is the ROA the same as or different from that of the industry?

2. (LO 3.3, 3.6) Consider the following table.

IQ Classifications in Educational Use

Wechsler, David. *Wechsler Adult Intelligence Scale-Third edition* Psychological Corporation, 1997

Classification	IQ Score	Percent Included	
		Theoretical Normal Curve	Actual Sample
Very Superior	130 and above	2.2	2.1
Superior	120–129	6.7	8.3
High Average*	110–119	16.1	16.1
Average	90–109	50.0	50.3
Low Average*	80–89	16.1	14.8
Borderline	70–79	6.7	6.5
Extremely Low* **	69 and below	2.2	1.9

Source: D. Wechsler, D. L. Coalson, and S. E. Raiford III, *WAIS-III: Wechsler Adult Intelligence Scale* (San Antonio, TX: Psychological Corporation, 1997).

   a. How is this IQ classification consistent with the *z*-score and confidence interval if the average population IQ is 100 and the standard deviation is 15?

   b. If your IQ is 130, the average population IQ is 100, and the standard deviation is 15, what is your *z*-score? How many standard deviations is that *z*-score above the mean? What does that suggest about your IQ compared to the population?

   c. If the average IQ of the population is 100 and the standard deviation is 15, what is the 95 percent confidence interval?

   d. How would you evaluate an IQ score of 65 using either a *z*-score or confidence interval?

3. (LO 3.8) Identify whether you expect a positive correlation, a negative correlation, or no correlation for the following scenarios.

   a. The relationship between time spent brushing your teeth and the number of cavities

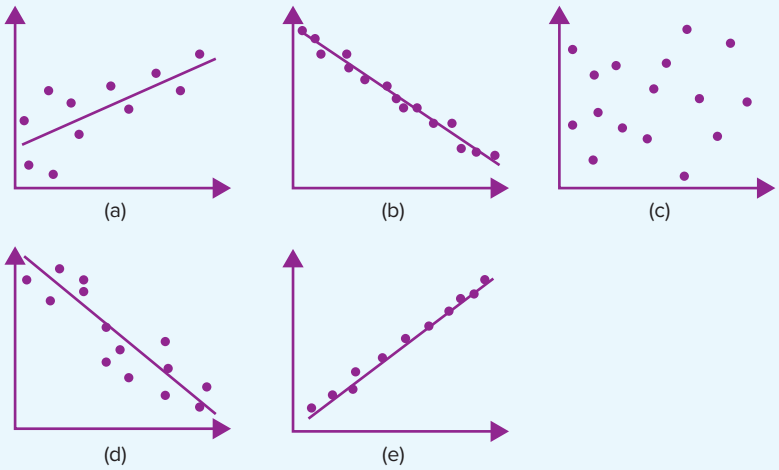
   b. The relationship between a car's age and auto repair costs

   c. The relationship between exercise time and calories burned

   d. The relationship between eye color and hair length

   e. The relationship between loan payments and the car owner's (remaining) level of debt owed on the car

4. (LO 3.8) Examine the scatterplots below. Identify which scatterplot exhibits:
   a. Strong positive correlation
   b. Positive correlation
   c. No correlation
   d. Negative correlation
   e. Strong negative correlation



5. (LO 3.8) A company wants to evaluate the relationship between sales revenue and advertising expenditures. A data analyst ran a regression analysis that generated the following output:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.892342619					
R Square	0.796275349					
Adjusted R Square	0.790283447					
Standard Error	10660.65663					
Observations	36					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	15103114544	1.51E+10	132.8919	2.71132E-13	
Residual	34	3864086391	1.14E+08			
Total	35	18967200936				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-806458.6655	77896.84187	-10.3529	4.76E-12	-964764.0947	-648153.24
Advertising Expense	152.2566144	13.20768996	11.52788	2.71E-13	125.415359	179.09787

Use the regression output to answer the following questions.

   a. What is the dependent variable? What is the independent variable?
   b. Is this an example of a simple regression or multiple regression?
   c. What does the Significance F result tell you about the regression?
   d. How many observations were used to generate the regression?
   e. What is the slope of the line?
   f. Is advertising expense positively or negatively related to sales revenue? Is the regression coefficient significant?

g. Evaluate the Adjusted R Square. What does it mean?

h. What is the line equation determined by regression?

i. If the company spends $10,000 on advertising expense, what is predicted sales revenue?

Use the following information to answer Problems 6 through 11.

Download the Skydio Drone dataset.xlsx. The data set contains sales transactions for the month of October for a company called Skydio that sells drones. SKUs (stock-keeping units) are the internal product codes used by Skydio to uniquely identify its products. You are going to use the data to evaluate the profitability of the company using gross (profit) margin and gross (profit) margin percentage. The data dictionary for the data follows.

Data Dictionary

Date:	Date of sales transaction
Invoice #:	Unique number identifying a sales transaction
Customer #:	Unique number identifying a customer
SKU:	Unique identifier for a product
Description:	Description of the SKU
Sales Price:	Sales price of the product (or SKU) (i.e., what the customer pays for the product)
Cost:	Cost of the product (or SKU) (i.e., what Skydio paid for the product)
Gross Margin:	How much Skydio profits from the item after subtracting the product's cost
Gross Margin Percentage:	Percentage of each dollar of sales revenue that Skydio retains as gross margin

Because of the way the data is structured (one item per row), an invoice may appear more than once in the data set. If a customer purchases one item, the invoice appears once in the data set. If an invoice includes more than one item, then the invoice appears more than once in the data set. For example, invoice 6007 has two rows in the data set, one for SK2-USB and one for SKY-CHG. Invoice 6054 has three rows in the data set, two rows for SKY2-CIK and one for SKY0-CHG.

To answer the questions, you must first add two columns in Excel to calculate the gross margin and gross margin percentage.

Column H Gross Margin = Sales Price – Cost

Column I Gross Margin Percentage = Gross Margin/Sales Price

6. (LO 3.1, 3.3, 3.4, 3.5) Let's analyze all of the sales transactions together. In Excel, calculate descriptive statistics for sales price, cost, gross margin, and gross margin percentage for all sales transactions. Use the output to answer the following questions:

   a. How many rows of data are in the data set?

   b. What is the average sales price, cost, and gross margin?

   c. How are the averages of sales price, cost, and gross margin related?

   d. Assuming the company incurs no other costs, how many dollars is the average profit on a transaction?

   e. On average, what percent profit does Skydio make on one item?

   f. How is the range calculated? Show an example using the numbers in the output.

   g. Is gross margin skewed? If so, is it positively skewed or negatively skewed? What does that mean?

7. (LO 3.3, 3.4) Let's analyze the sales transactions by SKU. Create a pivot table in Excel showing the average sales price, cost, gross margin, and gross margin percentage for each SKU. For ease of use, format all numbers with two decimals. Use the pivot table to answer the following questions:
   a. How many SKUs are there?
   b. Which SKU has the highest average selling price? (Hint: Sort the appropriate column.)
   c. Which SKU has the lowest average cost?
   d. Which SKU has the highest average gross margin?
   e. Which SKU has the highest average gross margin percentage?
   f. What is the Grand Total, or overall, average gross margin percentage?
   g. How can management use the information in the pivot table?
8. (LO 3.3, 3.4, 3.5) Let's analyze the transactions by product (SKU). Create a pivot table in Excel showing how many units were sold for each SKU. Use the pivot table to answer the following questions:
   a. Which product (SKU) sold the most units?
   b. Which product (SKU) sold the fewest units?
   c. Using PivotChart, create the recommended Cluster Column chart showing the number of products sold for each SKU. Is the chart a histogram or a bar chart? Explain your answer.
9. (LO 3.3, 3.4, 3.5) Let's analyze the transactions by customer. Create a pivot table in Excel showing how many invoices are associated with each customer. In addition, include a column showing the average gross profit margin percentage for each customer. Use the pivot table to answer the following questions:
   a. How many customers purchased items in October?
   b. Which customer had the highest number of sales transactions? (Hint: Use Count of Invoice #.)
   c. Which customer's sales transactions had the highest average gross margin percentage?
   d. How can management use the information in the pivot table?
10. (LO 3.1, 3.2, 3.3, 3.4, 3.5) Let's filter the data by customer. Make a copy of the data worksheet. Name the new worksheet tab Filter. Create a filter for customer number 3013. Copy the resulting data into row 235 (so you can create descriptive statistics). On the same worksheet, create descriptive statistics for customer 3013, placing the descriptive statistics in row 252. Answer the following questions about customer 3013.
    a. How many items did customer 3013 buy?
    b. Did customer 3013 buy more than one item in a sales transaction?
    c. What was the average gross margin percentage for customer 3013's sales transactions?
    d. Create a visualization showing the number of each product purchased by customer 3013 during October. Is the chart a histogram or a bar chart? Explain your answer.
    e. Why would management want to analyze sales by individual customers?
11. (LO 3.1, 3.6, 3.7) Management of Skydio thinks that the top 10 SKUs ranked by gross margin have a significantly different average gross margin than the bottom ten SKUs. Using the Data Analysis ToolPak, conduct a hypothesis test using a *t*-test assuming unequal variances to evaluate the premise. Hint: Create a pivot table for average gross margin by SKU; copy the numbers; sort and use the information to identify the 10 SKUs with the highest gross margins. Answer the following questions based on your hypothesis test.
    a. What is the null hypothesis? What is the alternative hypothesis?
    b. What is the average gross margin for the top 10 SKUs?
    c. What is the average gross margin for the bottom 10 SKUs?

d. What is the t Stat?

e. Assuming that the alpha is set at 0.05, would you reject or not reject the null hypothesis?

f. Assuming that the alpha is set at 0.01, would you reject or not reject the null hypothesis?

g. Is there a statistical difference between the top 10 and bottom 10 SKUs ranked by gross margin? Evaluate the results of the hypothesis test.



DATA VISUALIZATION

Data Visualization Exercise

1. (LO 3.1, 3.3, 3.5, 3.6, 3.7) Let's see how using a visualization can assist with data analysis. We will use a transformed version of the data in the file titled Skydio Drone dataset TRANSFORMED.xlsx. This is a transformed version of the data set used in Problems 6–11. This transformed file splits the first two weeks of data into two groups and lists each individual invoice amount. In the following three steps, assess if the average invoice for Week 1 is significantly different than the average invoice for Week 2. Complete the following steps:

   a. Calculate the mean for both sets of data.

   b. Create a bar chart for the two means. Why is a bar chart appropriate for these data?

   c. Based on the bar chart, do you think that the differences between the two groups are significantly different?

   d. Using the Data Analysis ToolPak, run a *t*-test assuming unequal variance.

      i. Select all of the data for Week 1 in the Variable 1 range and all of the data for Week 2 in the Variable 2 range (including the labels).

      ii. Place a checkmark next to Labels.

      iii. Set the hypothesized mean difference to 0.

      iv. Set the alpha to 0.05.

      v. Click OK to run the hypothesis test. The results will display automatically on a new sheet in your workbook.

   e. Interpret your results of both the bar chart and the *t*-tests (comparing the two-sample *p*-value to the alpha of 0.05).

   f. Would your results change if alpha were set to 0.10? Explain.

   g. Why do you think it is important not only to visually assess the differences between two data sets (as you did with the bar chart in this problem) but also to run a *t*-test to assess for statistical significance of the differences?

LAB 3.1 **Excel:** Using Excel Functions to Calculate Descriptive Statistics to Gain Insights About the Distribution of a Sales Data Set

LAB 3.2 **Excel:** Using the Excel Analysis ToolPak to Calculate Descriptive Statistics to Gain Insights About the Distribution of a Sales Data Set

LAB 3.2 **Tableau:** Calculating Descriptive Statistics to Gain Insights About the Distribution of a Sales Data Set

LAB 3.3 **Excel:** Performing a *t*-test for Difference in Means to Determine If the Differences Between In-Person and Online Sales Are Statistically Significant

LAB 3.4 **Excel:** Performing an ANOVA Test for Difference in Means to Determine If There Are Significant Differences Between the Average 4-Year Degree Completion Rate/SAT Average for Public, Private, and For-Profit Colleges

LAB 3.5 **Excel:** Deriving Cost Drivers for Activity-Based Costing (Regression Analysis)

Lab 3.1 Excel

Using Excel Functions to Calculate Descriptive Statistics to Gain Insights About the Distribution of a Sales Data Set

Keywords

Mean, Median, Standard Deviation, Range, Outlier, Box Plot, Histogram

Decision-Making Context

Retail companies need to monitor their sales to ensure they're meeting their goals. In this lab, we use Excel functions to calculate summary statistics for a company's total sales.

The Fruity Fruits Company sells exotic fruit from around the world to customers throughout the United States at retail outlets and online. It keeps a sales journal that records each sale and its related sales price and cost details.

Required

1. Calculate summary statistics using Excel functions.
2. Create visualizations showing the distribution of the sale transactions (box plot and histogram).

Specify the Question

What are the average, median, skewness, kurtosis, maximum, minimum, standard deviation, range, quartile 1, quartile 2, and interquartile range for the transaction amount? What does the distribution of the transaction amount look like?

Obtain the Data

This lab uses the data in Lab 3.1 Data.xlsx. This data set is a condensed version of Fruity Fruits' sales data that show the sales earned for each transaction and the transaction type (online or in-person). The data dictionary for this file is below.

Data Dictionary

Invoice #:	Unique number identifying a sales transaction
Transaction Amount:	Total sales revenue per transaction
Transaction Type:	Indicator of whether a transaction was in person or online

Analyze the Data

Analysis Task 1: Use Excel Functions to Calculate Summary Statistics.

1. In cell F2, calculate the average of the Transaction Amount. The function is =AVERAGE(array), so your function will look like Lab Exhibit 3.1.1. Click Enter.

Lab Exhibit 3.1.1

	A	B	C	D	E	F
1	Invoice #	Transaction Amount	Transaction Type			
2	3021	25.52	Online		Average	=AVERAGE(B2:B194)
3	3022	5.92	Online		Median	
4	3023	53.92	Online		Skewness	
5	3024	7.12	Online		Kurtosis	
6	3025	50.02	Online		Max. Value	
7	3026	50.02	Online		Min. Value	
8	3027	37.12	Online		Standard Deviation	
9	3028	2.52	Online		Range	

2. Using the same technique, calculate the following statistics:

In Cell . . .	Calculate the . . .	Using the function . . .
F3	Median	=MEDIAN(array)
F4	Skewness	=SKEW(array)
F5	Kurtosis	=KURT(array)
F6	Maximum value	=MAX(array)
F7	Minimum value	=MIN(array)
F8	Standard deviation	=STDEV.S(array)

Analysis Task 2: Calculate the Range and the Interquartile Range.

1. Calculate the range in cell F9: Subtract the value in cell F7 (the minimum value) from the value in cell F6 (the maximum value).
2. Calculate the interquartile range. This is a multi-step process: You will calculate the 25th percentile (quartile 1) and the 75th percentile (quartile 3), then find the difference between the two values.
   a. Calculate the 25th percentile in cell F10: The function is =QUARTILE.EXC(array,1). This function is slightly more advanced than the formulas used in the previous analysis task because it has two arguments. Not only does it include

the array selection, but you also must indicate which quartile you wish Excel to calculate for you. In this instance, we will use the number 1.

b. Calculate the 75th percentile in cell F11. This time you will use the same function you used to calculate the 25th percentile, but instead of indicating 1 for the quartile, you will indicate 3: =QUARTILE.EXC(array,3).

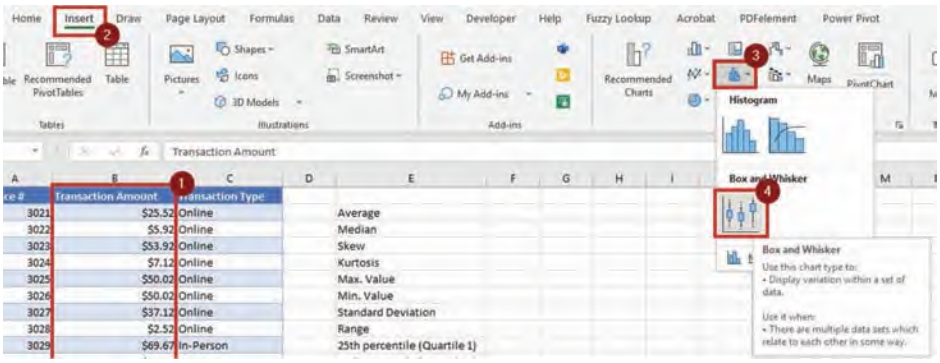
c. Calculate the interquartile range in cell F12: Subtract the 25th percentile from the 75th percentile.

Report the Results: Visualize the Distribution



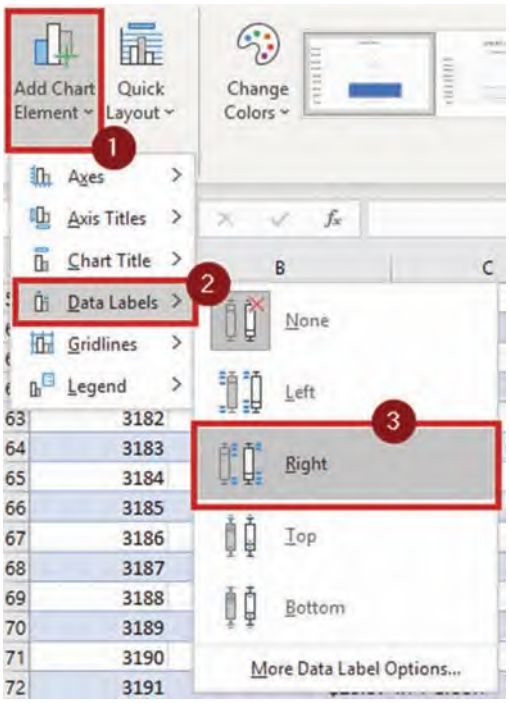
1. Insert a box plot to visualize the distribution and to gain insights into the range, the interquartile range, and the distinction between the median and the average. Highlight all of the data in the Transaction Amount column. From the Excel menu, select **Insert > Charts > Insert Statistic Chart > Box and Whisker**, as shown in Lab Exhibit 3.1.2.

Lab Exhibit 3.1.2



2. Add data labels to the chart. First, select the newly added Chart Tools Design tab. Then select **Add Chart Element > Data Labels > Right**, as shown in Lab Exhibit 3.1.3.

Lab Exhibit 3.1.3



3. Insert a histogram for another view of the distribution, including a clearer view of the skewness of the data. Highlight all of the data in the Transaction Amount column. From the Excel menu, select **Insert > Charts > Insert Statistic Chart > Histogram**.

Assessment

1. Take a screenshot of your summary statistics calculations, copy it into a Word document named "Lab 3.1 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the box plot, copy it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the histogram, copy it into the same Word document, and label the screenshot Submission 3.
2. Answer the questions in Connect and upload your Lab 3.1 Excel Submission.docx via Connect, if assigned.

Alternate Lab 3.1 Excel

Using Excel Functions to Calculate Descriptive Statistics to Gain Insights About the Distribution of a Sales Data Set

Apply the same steps in Lab 3.1 to the Alt Lab 3.1 Data.xlsx data set.

Required

1. Calculate summary statistics using Excel functions.
2. Create visualizations showing the distribution of the sale transactions (box plot and histogram).

Assessment

1. Take a screenshot of your summary statistics calculations, copy it into a Word document named "Alt Lab 3.1 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the box plot, copy it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the histogram, copy it into the same Word document, and label the screenshot Submission 3.
2. Answer the questions in Connect and upload your Alt Lab 3.1 Excel Submission.docx via Connect, if assigned.

Lab 3.2 Excel

Using the Excel Analysis ToolPak to Calculate Descriptive Statistics to Gain Insights About the Distribution of a Sales Data Set

Keywords

Mean, Median, Standard Deviation, Range, Outlier, Excel Analysis ToolPak, Box Plot

Decision-Making Context

Retail companies need to monitor their sales to ensure they are meeting their goals across different sales channels. In this lab, we use the Excel Analysis ToolPak to calculate summary statistics for a company's online and in-person sales to see how they compare. We then create a box plot to help visualize the distributions of the different types of sales.

The Fruity Fruits Company sells exotic fruit from around the world to customers throughout the United States at retail outlets and online. It keeps a sales journal that records each sale and its related sales price and cost details.

Required

1. Calculate summary statistics using the Excel Analysis ToolPak.
2. Create a box plot to show the distribution of the online and in-person sale transactions.

Specify the Question

What are the descriptive statistics for online and in-person sales? What does the distribution of the transaction amount look like? Based on the descriptive statistics and box plot, how do the online and in-person sales compare?

Obtain the Data

This lab uses the data in Lab 3.2 Data.xlsx. This data set is a condensed version of Fruity Fruits' sales data that show the sales earned for each transaction and the transaction type (online or in-person). The data dictionary for this file is below.

Data Dictionary

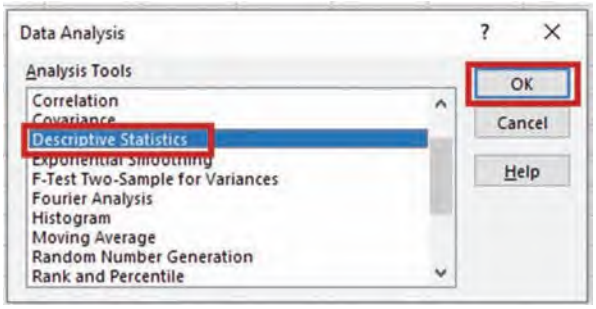
Invoice #:	Invoice number for each sale of fruit
Online:	Total sales revenue per transaction that was completed online
In-Person:	Total sales revenue per transaction that was completed in person

Analyze the Data

Analysis Task 1: Use Excel's Data Analysis ToolPak to Produce Summary Statistics.

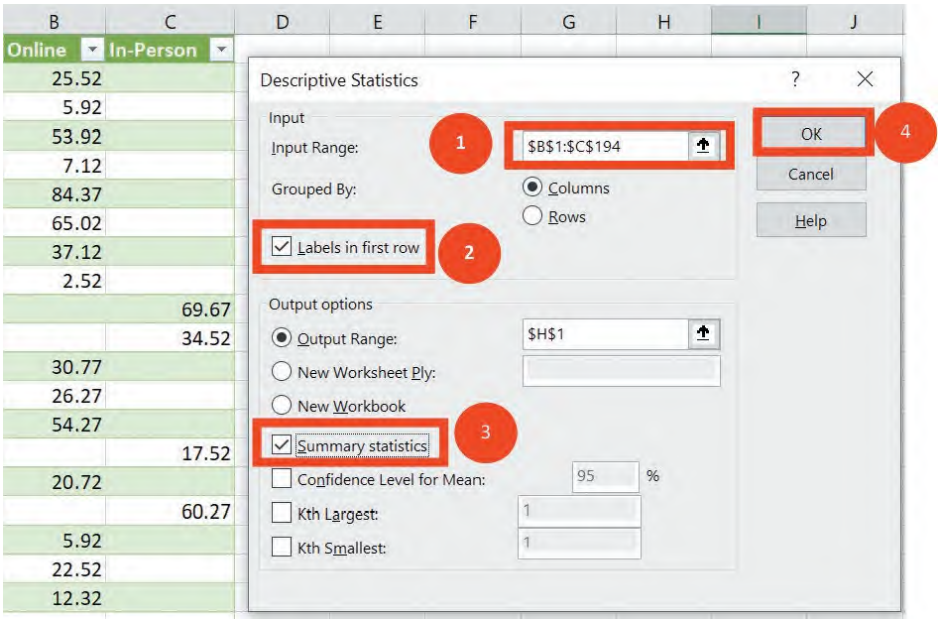
1. From the Excel menu, access the Data Analysis ToolPak from the Data tab. (If you do not have the Excel Analysis ToolPak downloaded, see Appendix E.)
2. Select Descriptive Statistics and click OK in the Data Analysis window, as shown in Lab Exhibit 3.2E.1.

Lab Exhibit 3.2E.1



3. Select the columns for Online and In-Person for the Input Range and check Labels in first row and Summary statistics. Click OK, as shown in Lab Exhibit 3.2E.2.

Lab Exhibit 3.2E.2



DATA VISUALIZATION

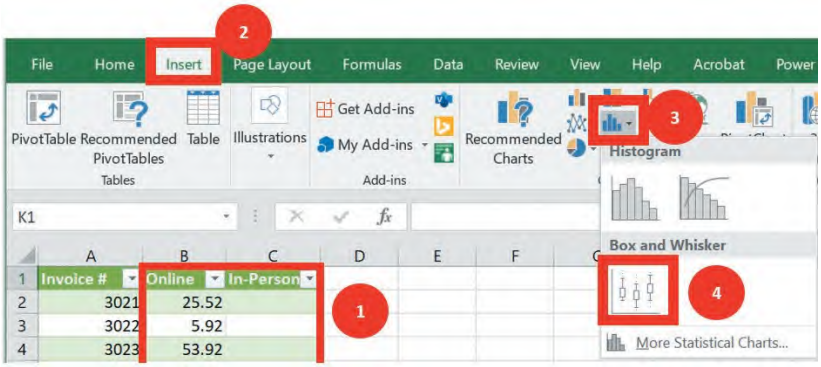
Report the Results: Create a Box-and-Whisker Plot to Visualize the Data

You have now produced the summary statistics and you can compare each statistic between online and in-person sales. Notice that the average and the median (measures of central tendency) for online sales are higher than the average and median for in-person sales. The standard deviation and the range (measures of dispersion) are also larger for online sales than for in-person sales. These measures indicate that although the average sale might be higher in an online environment, there is more variation in the sales amounts for online sales than in-person sales.

It can be easier to understand these distributions if we visualize the data. We will do so using box plots.

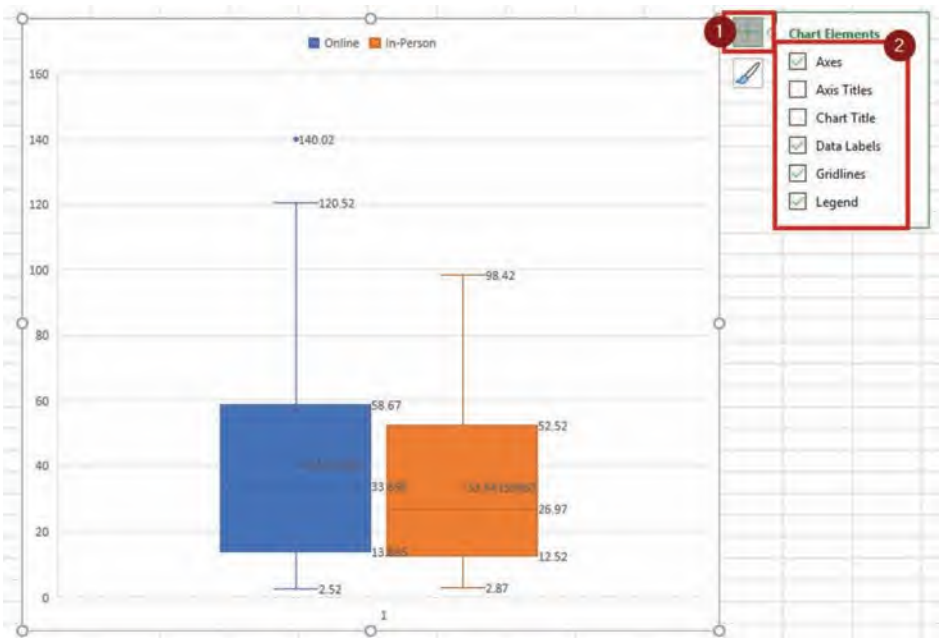
1. Highlight all of the data in the Online and In-Person columns. Then, from the Insert tab in the menu, select **Insert > Charts > Statistic Chart > Box and Whisker** (Lab Exhibit 3.2E.3).

Lab Exhibit 3.2E.3



2. Click the box plot so that the options button appears (a plus sign to the right of the chart area) and adjust the chart elements so that the Chart Title is no longer showing, but the Axes, Data Labels, Gridlines, and Legend are (Lab Exhibit 3.2E.4).

Lab Exhibit 3.2E.4



Adding the legend and the data labels to your box plots makes it easier to compare the values in the box plot to the values in your summary statistics output.

There are several items worth noting on the box plot.

- The observations we made concerning measures of central tendency and dispersion are easier to discern in this output. We can see that the range for the online sales plot is larger than for in-person sales, and we can see that the mean and median for the online sales are higher than those statistics for the in-person sales.
- Notice the online sales observation that falls above and outside the whiskers of the box plot is an outlier. This is not surprising given the large range for online sales.

We can report the results by sharing our findings that online sales seem to be higher than in-person sales based on the average sales amount. However, with the wide dispersion in values it may be worthwhile to dig further into the data to determine if the difference in average sales is due to chance or if it is statistically significant (which we will do in Lab 3.3).

Assessment

1. Take a screenshot of your summary statistics calculations, copy it into a Word document named "Lab 3.2 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the box plot, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Lab 3.2 Excel Submission.docx via Connect, if assigned.

Lab 3.2 Tableau

Calculating Descriptive Statistics to Gain Insights About the Distribution of a Sales Data Set

Keywords

Mean, Median, Standard Deviation, Range, Outlier, Box Plot

Decision-Making Context

Retail companies need to monitor their sales to ensure they are meeting their goals across different sales channels. In this lab, we use Tableau to calculate summary statistics for a company's online and in-person sales to see how they compare. We then create a box plot to help visualize the distributions of the different types of sales.

The Fruity Fruits Company sells exotic fruit from around the world to customers throughout the United States at retail outlets and online. It keeps a sales journal that records each sale and its related sales price and cost details.

Required

1. Visualize the dispersion of values using a box plot.
2. Add summary statistics.

Specify the Question

What are the descriptive statistics for online and in-person sales? What does the distribution of the transaction amount look like? Based on the descriptive statistics and box plot, how do the online and in-person sales compare?

Obtain the Data

This lab uses the data in Lab 3.2 Data.xlsx. This data set is a condensed version of Fruity Fruits' sales data that show the sales earned for each transaction and the transaction type (online or in-person). The data dictionary for this file is below.

Data Dictionary

Invoice #:	Invoice number for each sale of fruit
Online:	Total sales revenue per transaction that was completed online
In-Person:	Total sales revenue per transaction that was completed in person

Obtain Task 1: Load the Data into Tableau.

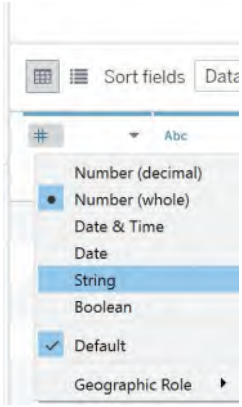
1. Open Tableau.
2. Under Connect, open Microsoft Excel.
3. In the Open window, browse to the Lab 3.2 Data.xlsx file. Click Open.

Obtain Task 2: Prepare the Data for Analysis.

1. Tableau defined Invoice # as a number (see the # symbol in the upper-left corner above Invoice #). Because Invoice # is an identifier/dimension, not a number used in computations, change Invoice # to a string (). To do this, click on the # above Invoice # and select String, as shown in Lab Exhibit 3.2T.1. Invoice # should now be identified as an "Abc" type of data. The data are ready to analyze.

Lab Exhibit 3.2T.1

Tableau Software, Inc. All rights reserved.



Analyze the Data

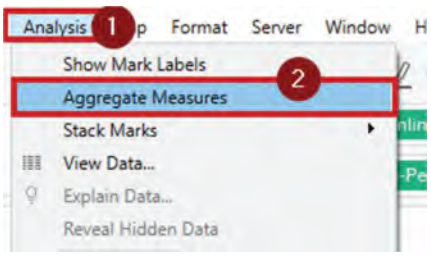
Analysis Task 1: Create a Summary Card.

Tableau uses a summary card to present descriptive statistics.

1. Click on Sheet 1 to go to a new Tableau sheet.
2. Add the In-Person and Online measures to the row and column shelves by double-clicking on each.
3. From the Analysis menu, remove the checkmark from Aggregate Measures by clicking it (Lab Exhibit 3.2T.2).

Lab Exhibit 3.2T.2

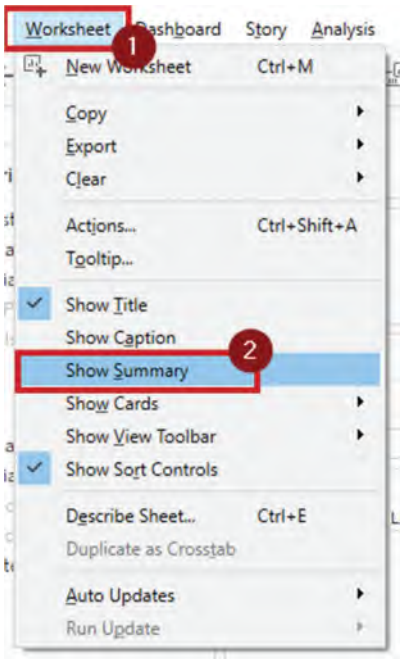
Tableau Software, Inc. All rights reserved.



4. To show Summary Statistics, select the Worksheet menu and click Show Summary to place a checkmark next to it and to view summary statistics (Lab Exhibit 3.2T.3).

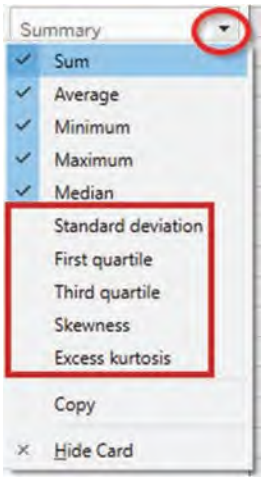
Lab Exhibit 3.2T.3

Tableau Software, Inc. All rights reserved.



The summary card shows up on the right of your Tableau window. (You may need to hide the Show Me menu in order to see the summary card.) The summary card defaults to showing only Sum, Average, Minimum, Maximum, and Median.

If you would like to see other values such as standard deviation and range, you can click the drop-down menu next to Summary and select all of the unchecked statistics to place a checkmark next to them and view them (Lab Exhibit 3.2T.4).



Lab Exhibit 3.2T.4

Tableau Software, Inc. All rights reserved.

Report the Results: Create a Box-and-Whisker Plot to Visualize the Data



DATA VISUALIZATION

Let's create a box plot to clearly show how the data are dispersed for in-person and online sales.

1. Click Sheet 2.
2. Add the In-Person and Online measures to the row and column shelves by double-clicking each.
3. From the analysis menus, remove the checkmark from Aggregate measures.
4. Change the default visualization to a box plot by selecting box-and-whisker plots from the Show Me menu (Lab Exhibit 3.2T.5).
   From this visualization we can see how the distribution differs between in-person and online sales. We can learn even more by adding in summary statistics and labels.



Lab Exhibit 3.2T.5

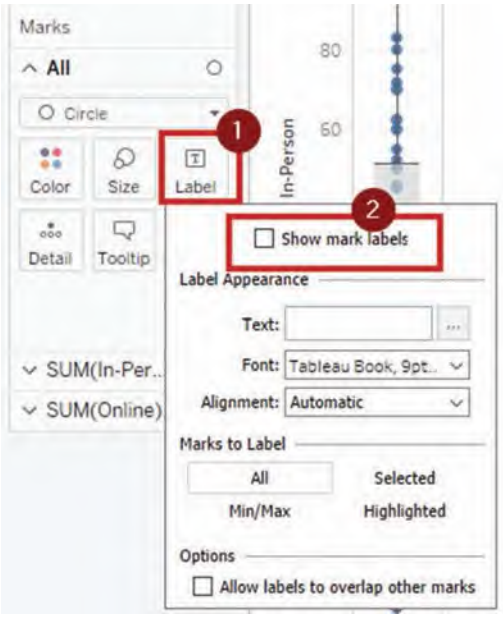
Tableau Software, Inc. All rights reserved.

5. In the Marks shelf, select Label and check Show mark labels (Lab Exhibit 3.2T.6). The maximum value for each distribution will appear. If you would like more detail, you can hover your mouse over any observation in the box plot to see the observation.

We can report the results by sharing our findings that online sales seem to be higher than in-person sales based on the average sales amount. However, online sales has a larger

Lab Exhibit 3.2T.6

Tableau Software, Inc. All rights reserved.



range of values than in-person sales. It may be worthwhile to dig further into the data to determine if the difference in average sales is due to chance or if it is statistically significant (which we will do in Lab 3.3).

Assessment

1. Take a screenshot of your summary card statistics calculations, copy it into a Word document named "Lab 3.2 Tableau Submission.docx", and label the screenshot Submission 1. Take a screenshot of the box plot, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Lab 3.2 Tableau Submission.docx via Connect, if assigned.

Alternate Lab 3.2

Calculating Descriptive Statistics to Gain Insights About the Distribution of a Sales Data Set

Using either Excel or Tableau, apply the same steps in Lab 3.2 to the Alt Lab 3.2 Data.xlsx dataset.

Assessment

1. Take a screenshot of your summary card statistics calculations, copy it into a Word document named "Alt Lab 3.2 Submission.docx", and label the screenshot Submission 1. Take a screenshot of the box plot, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Alt Lab 3.2 Submission.docx via Connect, if assigned.

Lab 3.3 Excel

Performing a *t*-test for Difference in Means to Determine If the Differences Between In-Person and Online Sales Are Statistically Significant

Keywords

Hypothesis Test, Statistical Significance, *p*-Value, Alpha, Box Plot

Lab Insight

Retail companies need to monitor their sales to ensure they're meeting their goals across different categories, time periods, and sales methods. In this lab, we calculate summary statistics for a company's online sales and its in-person sales to determine if there is a difference. If the company finds a statistically significant difference in one method of sales over the other, it may take action to increase sales in the lower-performing method, perhaps by putting together a marketing campaign.

While it is reasonable to assume that the average sales transaction amount for in-person sales is not identical to the average sales transaction amount for online sales, it's important to determine if that difference is statistically significant before taking any actions. In other words, we need to ask: Is the difference in average sales likely just due to chance? Or is the difference large enough that we can assume it is truly (at least in part) due to the different sales channel? In this lab we test for significant differences between online and in-person sales for Fruity Fruits Company.

The Fruity Fruits Company sells exotic fruit from around the world to customers throughout the United States at retail outlets and online. It keeps a sales journal that records each sale and its related sales price and cost details.

Required

1. Construct hypotheses regarding the differences in means.
2. Perform a *t*-test for differences in means using the Data Analysis ToolPak.
3. Interpret the results by comparing the *p*-value to alpha.
4. Create a box-and-whisker plot to see the distribution.

Specify the Question

Is there a significant difference between average sales revenue per transaction for in-person and online sales?

Obtain the Data

This lab uses the data in Lab 3.3 Data.xlsx. This data set is a condensed version of Fruity Fruits' sales data that show the sales earned for each transaction and the transaction type (online or in-person). The data dictionary for this file is below.

Data Dictionary

Invoice #:	Invoice number for each sale of fruit
Online:	Total sales revenue per transaction that was completed online
In-Person:	Total sales revenue per transaction that was completed in person

Analyze the Data

Analysis Task 1: Determine the Hypotheses.

We will use the standard pattern set forth in this chapter to specify the null and alternative hypotheses.

H_o: There is *not* a significant difference between the average sales revenue for online transactions and the average sales revenue for in-person transactions.

H_A: There *is* a significant difference between the average sales revenue for online transactions and the average sales revenue for in-person transactions.

Based on the alternative hypothesis, we will be conducting a two-tailed hypothesis test as no direction is specified.

Analysis Task 2: Set the Statistical Significance (the Alpha).

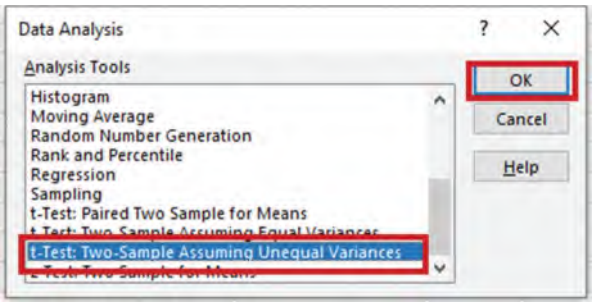
We will set the significance at 0.05.

Analysis Task 3: Calculate the Test Statistic.

To conduct the hypothesis test, we use the Excel Analysis ToolPak (see Appendix E for instructions on how to install). Make sure the ToolPak is installed by selecting the Data tab. If Data Analysis is shown in the Analyze group, you are ready to go. Conduct the hypothesis test by following the steps below.

1. From the Excel menu, select **Data > Analyze > Data Analysis.**
2. Select t-Test: Two-Sample Assuming Unequal Variances and click OK (Lab Exhibit 3.3.1).

Lab Exhibit 3.3.1



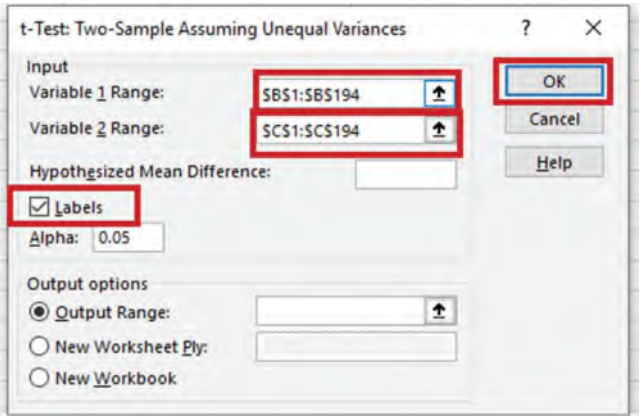
3. In the t-test: Two Sample Assuming Unequal Variances dialog box, you have options to input several fields. Here are explanations of each field, as well as what to input (described below and also shown in Lab Exhibit 3.3.2):

   a. Variable 1 Range: This is all of the data (including the header description) for the first set of transactions that you are comparing. Select all of the transactions from the Online column for this input. The range should be $B$1:$B$194.
   b. Variable 2 Range: This is all of the data (including the header description) for the second set of transactions that you are comparing. Select all of the transactions from the In-Person column for this input. The range should be $C$1:$C$194.
   c. Hypothesized Mean Difference: When testing for differences between means, Excel interprets the default value of a blank as 0. If, however, you wanted to analyze your data for a mean difference greater than 0, you could change this input. For this lab, we will leave it blank so that it defaults to 0.

d. Labels: We included the headers for Online and In-Person for the Variable 1 and 2 Range inputs, so check the box labeled Labels.
e. Alpha: This defaults to 5% (0.05), which matches the selected significance level. If you would like a more restrictive alpha such as 1% or a less restrictive alpha such as 10%, you can adjust it in this input. In this lab, we will leave alpha at the default of 5%.
f. Output options. If you want to put the hypothesis test results on the same sheet as the data, click on Output Range and indicate where you want the data to appear, for example, $H$1. Alternatively, if you want to put the hypothesis test results on a new worksheet, click on New Worksheet Ply. For this lab you can choose which option you would like.

Once you have finished your inputs, click OK to run the hypothesis test.

Lab Exhibit 3.3.2



Analysis Task 4: Reject or Fail to Reject the Null Hypothesis Using the *p*-Value.

1. **Interpret the results of a two-tailed test:**
   Based on the alternative hypothesis, we are using a two-tailed test because we did not specify a specific direction. Based on the Excel output, the t Stat is 1.60 with a two-tailed *p*-value of 0.112203. To determine whether we "reject" or "fail to reject" the null hypothesis, we compare that *p*-value to alpha.
   We set alpha as 5%, or 0.05, and 0.112203 is greater than .05. This means we fail to reject the null hypothesis.
   Remember that rejecting the null hypothesis supports the notion that the averages are significantly different. In contrast, when we fail to reject the null hypothesis, as in this case, we recognize that the differences in the averages are not significant. If management is happy that the average sales revenue from a transaction is significantly the same regardless of the sales channel (online or in person), no action will be taken.
2. **Interpret the results of a one-tailed test:**
   What if we had specified the following alternative hypothesis:

   H_A: The average sales revenue for an online transaction is significantly *greater* than the average sales revenue for an in-person transaction.

   This alternative hypothesis requires a one-tailed test because it specifies a direction. Comparing the one-tailed *p*-value of 0.056101 to an alpha of 0.05, we still fail to reject the null hypothesis, indicating that the differences in means is due to chance.

Notice that the *p*-value is much closer to being significant in this one-tailed case than it was in the two-tailed test. But it is still insignificant at a 5% level.
If we set alpha at 10% to determine statistical significance, we would reject the null hypothesis for the one-tailed test. If we wanted average sales revenue for a transaction to be the same in both channels, the company might decide to take action, perhaps by changing its marketing approach for in-person sales. This example shows the importance of choosing an appropriate alpha at the beginning of your business analysis project.

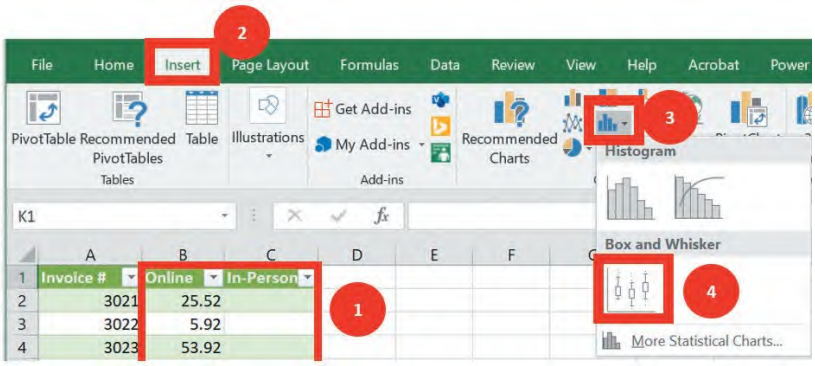

DATA VISUALIZATION

Report the Results: Create a Box-and-Whisker Plot to Visualize the Data

We can report the results by reporting our findings that while online stores seem to be more successful than in-person sales based on the average sales amount, the difference is not statistically significant and is likely due to chance. When we present the findings to management we may want to include a visualization. A box-and-whisker plot may be helpful in showing distributions. Follow the directions below to create a box-and-whisker plot.

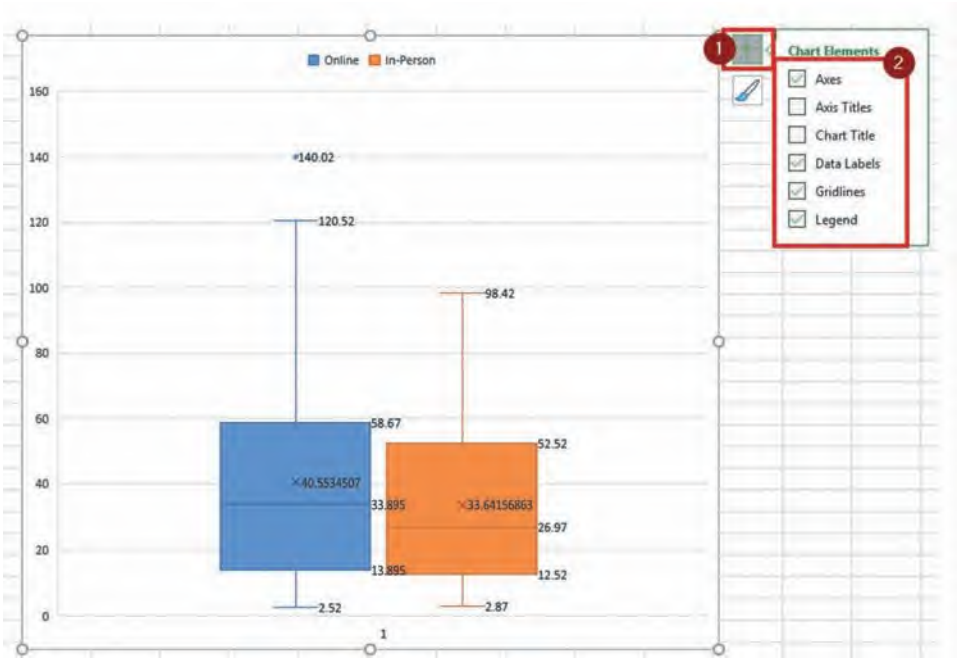
1. Highlight all of the data in the Online and In-Person columns. Then, from the Insert tab in the menu, select **Insert > Charts > Statistic Chart > Box and Whisker** (Lab Exhibit 3.3.3).

Lab Exhibit 3.3.3



2. Click the box plot so that the options button appears (a plus sign to the right of the chart area) and adjust the chart elements so that the Chart Title is no longer showing, but the Axes, Data Labels, Gridlines, and Legend are (Lab Exhibit 3.3.4).

Lab Exhibit 3.3.4



Assessment

1. Take a screenshot of your hypothesis test, copy it into a Word document named "Lab 3.3 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the box plot, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Lab 3.3 Excel Submission.docx via Connect, if assigned.

Alternate Lab 3.3

Performing a *t*-test for Difference in Means to Determine If the Differences Between In-Person and Online Sales Are Statistically Significant

Apply the same steps in Lab 3.3 to the Alt Lab 3.3 Data.xlsx data set.

Assessment

1. Take a screenshot of your hypothesis test, copy it into a Word document named "Alt Lab 3.3 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the box plot, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Alt Lab 3.3 Excel Submission.docx via Connect, if assigned.

Lab 3.4 Excel

Performing an ANOVA Test for Difference in Means to Determine If There Are Significant Differences Between the Average 4-Year Degree Completion Rate/SAT Average for Public, Private, and For-Profit Colleges

Keywords

ANOVA, *p*-Value, Null Hypothesis, Alternative Hypothesis

Lab Insight

The College Scorecard data set is provided by the U.S. Department of Education. It provides federal financial aid and earnings information, insights into the performance of schools eligible to receive federal financial aid, and the outcomes of students at these schools. We have selected a subset of the data to focus just on the average completion rates and the SAT averages (for Alternate Lab 3.4) for public, private, and for-profit colleges.

As a college student, you decide which type of college you wanted to attend. The three possible categories are public, private, and for-profit, each with a different average completion rate. However, are those average completion rates *significantly* different from each other? In this lab, we use the ANOVA test to determine if the averages are significantly different.

Required

1. Construct hypotheses regarding the differences in means.
2. Perform an ANOVA test for differences in the group means using the Data Analysis ToolPak.
3. Interpret the results by comparing the *p*-value to alpha.
4. Create a box-and-whisker plot to see the distribution.

Specify the Question

Is there a significant difference between the average completion rates at different categories of colleges?

Obtain the Data

This lab uses the data in Lab 3.4 Data.xlsx. This data set is a sample of the data available from the U.S. Department of Education. Each cell contains the graduation or completion rate for a college. The column headings identify the type of college (public, private nonprofit, private for-profit). The data dictionary for this file is below.

Data Dictionary

Public:	Completion rates at public colleges
Private Nonprofit:	Completion rates at private nonprofit colleges
Private for-Profit:	Completion rates at private for-profit colleges

Analyze the Data

Analysis Task 1: Determine the Hypotheses.

The two hypotheses in this ANOVA analysis are:

H_o: There is no significant difference among the average completion rates for the three college categories. (Or, all average completion rates are equal.)

H_A: There is an overall significant difference among the average completion rates for the three college categories. (Or, all average completion rates are not equal.)

Analysis Task 2: Set the Statistical Significance (the Alpha).

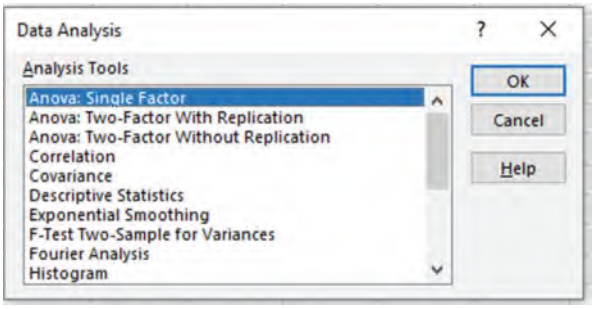
We will set the significance at 0.05.

Analysis Task 3: Calculate the Test Statistic.

To conduct the ANOVA, we use the Excel Analysis ToolPak (see Appendix E for instructions on how to install). Make sure the ToolPak is installed by selecting the Data tab. If Data Analysis is shown in the Analyze group, you are ready to go. Conduct the ANOVA by following the steps below.

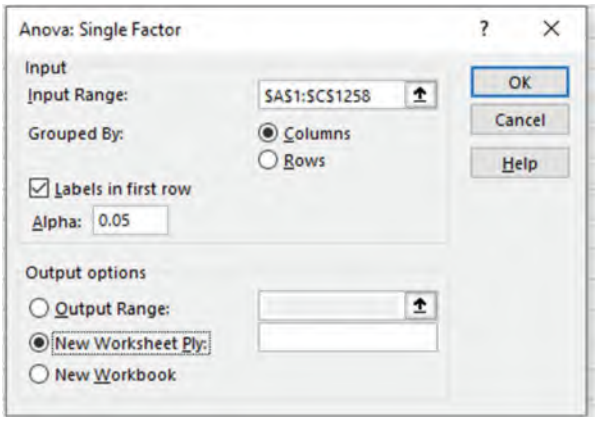
1. From the menu, select **Data > Analyze > Data Analysis.** A dialog box will open. Select ANOVA: Single Factor (Lab Exhibit 3.4.1).

Lab Exhibit 3.4.1



2. Select the full set of data for the Input Range, check the box next to Labels in the first row, accept the default value for Alpha of 0.05, and click OK (Lab Exhibit 3.4.2).

Lab Exhibit 3.4.2



Lab Exhibit 3.4.3 shows the results of the ANOVA test. The SUMMARY table reports the number of observations (the Count), as well as the total, average, and variance for each college category. The statistic of interest is the mean or average. From smallest to largest, private for-profit colleges have a graduation or completion rate of 38 percent; public colleges have a completion rate of 46 percent; and private nonprofit colleges have a 54 percent completion rate. So, private nonprofit colleges have the highest graduation rates but also the highest variance. The ANOVA test examines whether the differences in the group means are statistically significant.

Lab Exhibit 3.4.3

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Public	668	304.4049	0.455696	0.032313
Private nonprofit	1257	682.7093	0.543126	0.04245
Private for Profit	521	198.6982	0.381379	0.038351

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	10.41465	2	5.207326	134.1748	0	2.999409
Within Groups	94.81287	2443	0.03881			
Total	105.2275	2445				

Analysis Task 4: Reject or Fail to Reject the Null Hypothesis Using the *p*-Value.

The ANOVA table presents statistics. The most important statistic for our ANOVA test is the *p*-value, which we need to compare to the predefined alpha or significance level, which we set at 0.05. In this instance, our *p*-value is extremely small—it is 0! Because 0 is less than 0.05, we can reject our null hypothesis and conclude that not all of the group means are equal. In other words, there are significant differences among the average completion rates for the three different categories of colleges. This means that the three types of colleges do not all have the same average completion rate.



DATA VISUALIZATION

Report the Results: Graph Confidence Intervals.

To help present our analysis, let's create a graph showing the confidence intervals for the three categories of colleges. Recall that a **confidence interval** provides a range of numbers (an interval) around the point estimate constructed so that we are very confident that the interval contains the true population parameter at a certain level of confidence, such as 90% or 95%.

Also recall from the chapter that a confidence interval is equal to the point estimate ± the margin of error.

Margin of error = 1.96 * (standard deviation / the square root of *n*).

We will calculate the margin of error to determine the confidence intervals. We are using 1.96 because it corresponds to a 95 percent confidence interval.

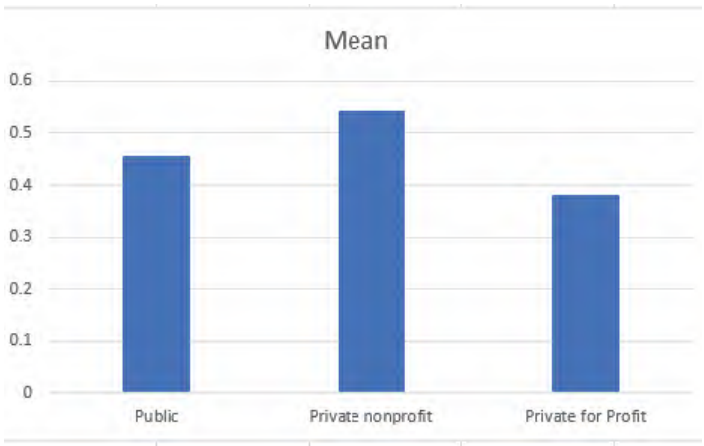
1. In the ANOVA tab of the worksheet, set up the following table shown in Lab Exhibit 3.4.4 in L1 through N4. The formula for Margin of Error in cell N2 = 1.96 *(E5^0.5)/(B5^0.5). E5 is the variance, so the square root (^0.5) of the variance is the standard deviation. B5 is the count or number of observations (*n*). Copy the formula down the column. The spreadsheet should look like Lab Exhibit 3.4.4.

Lab Exhibit 3.4.4

	L	M	N
1	Category	Mean	Margin of Error
2	Public	0.455696	0.013631851
3	Private nonprofit	0.543126	0.011390152
4	Private for Profit	0.381379	0.016816109

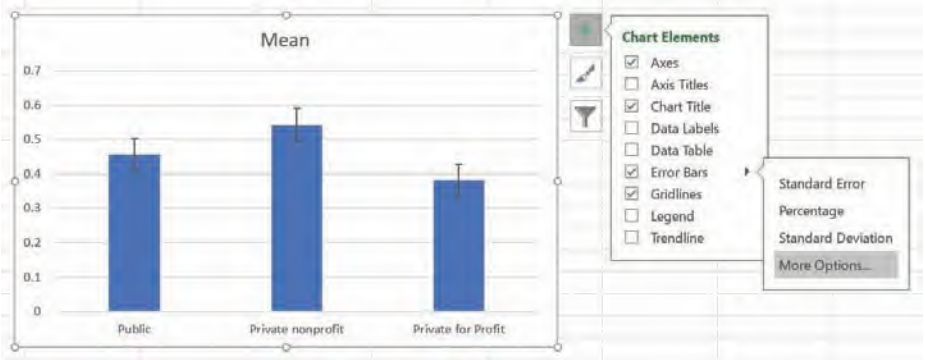
2. Highlight L1 through M4.
3. From the menu, select **Insert > Charts > Recommended Chart > Clustered Column**. The bar chart should look like Lab Exhibit 3.4.5.

Lab Exhibit 3.4.5



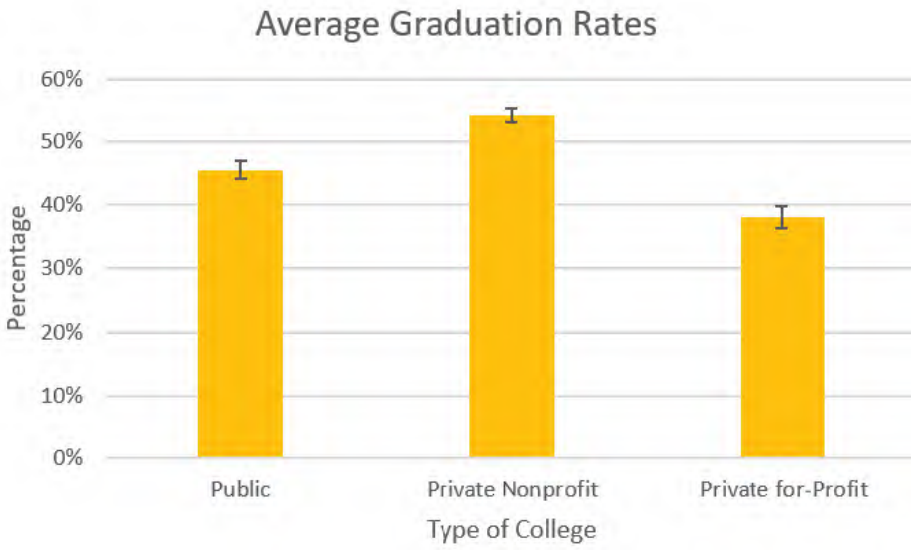
4. Click on the chart. The + sign should appear on the upper-right corner.
5. Click on Error Bars. Then click the expansion arrow by Error Bars and select More Options, as shown in Lab Exhibit 3.4.6.

Lab Exhibit 3.4.6



6. In the Format Error Bars dialog box, scroll down and select Custom. Click Specify Value.
7. In the new window that appears, insert N2:N4 for both the positive and negative error value. Click OK.
8. Change the bar colors to make the bar chart easier to read by right-clicking each bar and selecting a lighter color.
9. Change the title of the chart to "Average Graduation Rates".
10. Let's show the Y-axis as a percentage. Right-click on the Y-axis. Select Format Axis. In the Format Axis dialog box, scroll down to Number. Set category to Percentage with zero decimals.
11. Let's add a label for the X-axis and Y-axis. Select **Chart Tools Design > Add Chart Element > Add Axis Titles** to add both X-axis (primary horizontal) and Y-axis (primary vertical). Label the X-axis "Type of College". Label the Y-axis "Percentage".

Lab Exhibit 3.4.7



Lab Exhibit 3.4.7 shows the resulting bar chart with confidence intervals. The top of each bar chart represents the mean graduation rate for each sample. The whisker is the 95% confidence interval around the mean. For example, for the first bar, Public colleges, the confidence interval represented by the whisker ranges from 0.44206 to 0.46933 (the mean +/− the margin of error). The chart shows that the confidence intervals are very small, indicating that the true population mean should be very close to the sample mean.

Assessment

1. Take a screenshot of your ANOVA test, copy it into a Word document named "Lab 3.4 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the confidence interval, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Lab 3.4 Excel Submission.docx via Connect, if assigned.

Alternate Lab 3.4

Excel: Performing an ANOVA Test for Difference in Means to Determine If There Are Significant Differences Between the Average 4-Year Degree Completion Rate/ SAT Average for Public, Private, and For-Profit Colleges

Apply the same steps in Lab 3.4 to the Alt Lab 3.4 Data.xlsx data set.

1. Take a screenshot of your ANOVA test, copy it into a Word document named "Alt Lab 3.4 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the confidence interval, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Alt Lab 3.4 Excel Submission.docx via Connect, if assigned.

Lab 3.5 Excel

Deriving Cost Drivers for Activity-Based Costing (Regression Analysis)

Keywords

Cost drivers, Overhead Allocation, Activity-Based Costing, Regression

Decision-Making Context

The price of a product is often based on the cost of a product. The cost of a product can be expressed by the following formula:

Product Cost = Direct Material Costs + Direct Labor Costs + Factory Overhead Costs

Direct costs are costs that can be tracked directly to a product. Examples include the raw materials to make a product and assembly line labor costs to produce the product. In addition to direct costs, making a product also requires overhead costs. **Overhead costs** are indirect costs that are generally hard to trace to an individual product. Overhead costs include taxes, rent, insurance, business permits, maintenance and repair of equipment, utilities (electricity, telephone), and other factory costs.

Direct costs are generally easy to calculate because the company knows how much raw material is required to make a product as well as the direct labor cost to assemble a product (required hours times the hourly labor rate). Indirect costs, however, cannot be directly traced to a product. Thus, indirect costs are *allocated* to a product. It is important to allocate those costs to adequately price and evaluate the cost of producing a product for sale.

Many corporations use activity-based costing to allocate overhead costs. **Activity-based costing (ABC)** assumes that the best way to assign indirect costs to products is based on the products' demand for resource-consuming activities, which we call **cost drivers**. In some sense, we are trying to make indirect costs come as close to direct costs as we can so that we can determine if we would make different decisions based on that cost allocation.

Bob's Delivery, a fictitious company, delivers bulk paper products to its customers. The business incurs various indirect costs, including car maintenance, tires, oil changes, gasoline, car insurance, and car property tax, that it would like to allocate to each delivery. If managers can more accurately allocate these indirect or overhead costs, then the business can more accurately charge its customers for their deliveries. Managers would like to estimate cost drivers as part of an activity-based costing system. In this lab, we use regression analysis to derive the cost drivers used in ABC, and then we derive a costing formula using the cost drivers to allocate overhead for Bob's Delivery.

Required

1. Using regression analysis and choosing from several potential cost drivers, derive the cost drivers that are significant predictors of overall overhead costs and can be used to help allocate overhead. Set alpha to 5%, or 0.05.
2. Derive the formula for allocating overhead (the regression line equation).
3. Compare estimated overhead per delivery to actual overhead per delivery using a scatterplot.

Specify the Question

How can we use regression analysis to find cost drivers and derive a formula for allocating overhead?

Obtain the Data

Bob's Delivery does an intense analysis of overhead costs. Once the managers identify the drivers of overhead costs, they develop a formula to allocate the overhead based on those cost drivers. The analysis calculates total overhead costs over 122 delivery days and considers three potential cost drivers that the company believes may be associated with the overhead costs. These potential cost drivers are the number of miles driven, the total delivery time, and the weight. This lab uses the data in Lab 3.5 Data.xlsx. This data set contains information for the potential cost drivers. The data dictionary for this file is below.

Data Dictionary

Delivery Day:	Indicates delivery day (total cost was estimated from days 1 to 122)
Total Overhead Cost:	The total cost for that delivery day
Deliveries:	The number of deliveries made that day
Miles:	The number of miles to make all deliveries that day
Delivery Time (Minutes):	The delivery time it takes to make all deliveries that day
Weight of Delivery (lbs):	Combined weight (in pounds) of all deliveries made that day

Analyze the Data

Analysis Task 1: Set Up the Spreadsheet.

1. To select the entire data set, place your cursor anywhere in the data and click Ctrl+A.
2. To format the data as a table, click **Home > Styles > Format** and choose any style. Check the box next to My table has headers and click OK.
3. We begin by adding titles. Add the following titles to these cells as shown in Lab Exhibit 3.5.1.

Lab Exhibit 3.5.1

Cell	Title
G1	Overhead Cost per Delivery
H1	Miles per Delivery
I1	Time per Delivery
J1	Weight per Delivery

4. We need to enter formulas to calculate the cost, miles, time, and weight for each delivery made during the day. Here is a description of the required formulas, followed by the actual formulas in Lab Exhibit 3.5.2.
   - In cell G2, calculate the average overhead cost per delivery by taking total overhead cost (in cell B2) and dividing by the number of deliveries (in cell C2).
   - In cell H2, calculate the average miles per delivery by taking total miles (in cell D2) and dividing by the number of deliveries (in cell C2).
   - In cell I2, calculate the average time per delivery by taking total delivery time (in cell E2) and dividing by the number of deliveries (in cell C2).
   - In cell J2, calculate average weight per delivery by taking total weight of delivery (in cell F2) and dividing by the number of deliveries (in cell C2).

For all of these cells, enter the formulas shown in Lab Exhibit 3.5.2 in cells G2:J2. Because this is a table, the formulas automatically fill down the column, with all numbers shown with two decimals.

Lab Exhibit 3.5.2

	G	H	I	J
1	Overhead Cost per Delivery	Miles per Delivery	Time per Delivery	Weight per Delivery
2	=B2/C2	=D2/C2	=E2/C2	=F2/C2

After entering these formulas, your cells should appear as shown in Lab Exhibit 3.5.3.

Lab Exhibit 3.5.3

	G	H	I	J
1	Overhead Cost per Delivery	Miles per Delivery	Time per Delivery	Weight per Delivery
2	72.67	22.38	23.50	113.13
3	56.50	12.00	21.00	519.00
4	60.90	22.25	18.25	476.25
5	70.46	19.31	27.50	110.06

Analysis Task 2: Run a Regression Analysis to Identify Relevant Cost Drivers.

We are now ready to perform the analysis. Regression analysis is a way of using mathematics and statistics to measure the relationship between one output variable and one or more input variables. Regression analysis helps us answer the questions: Which factors matter the most in predicting the outcome? Which factors can we ignore?

In this case, we are trying to figure out which cost drivers are associated with the outcome, or total overhead costs.

We can represent this situation as an algebraic equation in which y is the dependent variable and x represents the independent variables, where $y = f(x)$. In other words, we will run an equation where the dependent variable y, (total) overhead cost per delivery, is a function of independent variables. In this lab, we consider the following three potential independent variables or cost drivers:

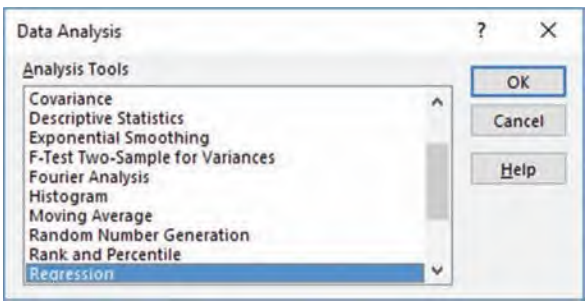
- **Miles per delivery:** The number of miles for each delivery
- **Time per delivery (minutes):** The time (in minutes) it takes to make each delivery
- **Weight per delivery:** The weight (in pounds) of each delivery

Using regression analysis, we can assess which cost drivers impact the overhead cost per delivery, as well as the size of that impact, by looking at the cost drivers that are statistically related to overhead cost. We can then derive a formula to allocate overhead to future deliveries.

To run the regression, we use the Excel Analysis ToolPak (see Appendix E for instructions on how to install). Make sure the ToolPak is installed by selecting the Data tab. If Data Analysis is shown in the Analyze group, you are ready to go. Run the regression by following the steps below.

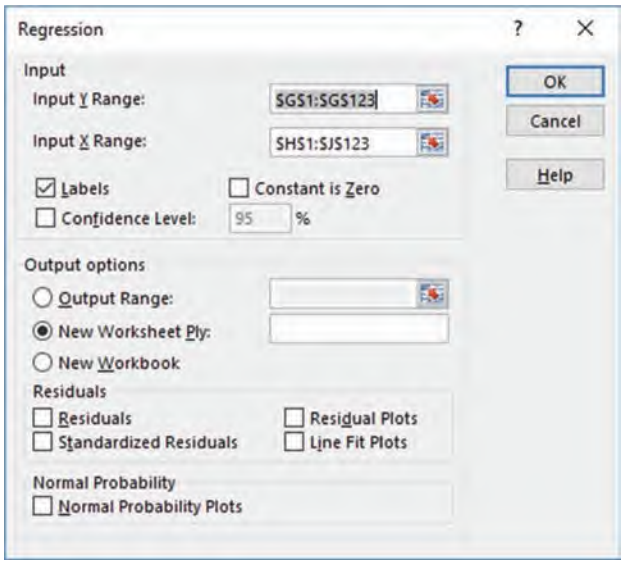
1. To run the regression, select **Data > Analyze > Data Analysis.** A dialog box will open. Select Regression, as shown in Lab Exhibit 3.5.4.

Lab Exhibit 3.5.4



2. A dialog box will open. We need to select the dependent variable (or Y Range). Our dependent variable is overhead cost per delivery, so enter $G$1:$G$123 as the Input Y Range. Note: You must enter the exact range. If you don't, Excel will generate an error.
3. Next, we need to select our independent variable (or X Range). Our cost drivers (miles per delivery, time per delivery, and weight per delivery) are our independent variables, so enter "$H$1:$J$123" in the Input X Range. Note: You must enter the exact range. If you do not, Excel will generate an error.
4. Because we highlighted the labels in the data range, click the box by "Labels". Also select New Worksheet Ply to get the results in a new worksheet. The completed Regression dialog box should look like Lab Exhibit 3.5.5.

Lab Exhibit 3.5.5



5. Now click OK to generate the regression results shown in Lab Exhibit 3.5.6.

Lab Exhibit 3.5.6

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.896794					
5	R Square	0.804239					
6	Adjusted R Square	0.799262					
7	Standard Error	4.494111					
8	Observations	122					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	9791.057	3263.686	161.5924	1.2764E-41	
13	Residual	118	2383.249	20.19703			
14	Total	121	12174.31				
15							
16		*Coefficients*	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	15.67445	2.968641	5.280007	5.97E-07	9.795730433	21.55317
18	Miles per Delivery	0.4548	0.096881	4.694405	7.27E-06	0.262948651	0.646651
19	Time per Delivery	1.880461	0.117657	15.98258	2.67E-31	1.64746813	2.113454
20	Weight per Delivery	-0.00226	0.002853	-0.79228	0.42979	-0.007909847	0.003389

There are many things to note about the output from the regression. The first is that the regression did better than chance at predicting overhead costs. We draw that conclusion by

noting that the "Significance F" result is very small, suggesting there is an extremely small probability that the overhead costs can be explained by a model that has no cost drivers (the independent variables). This result suggests that the predictive model with the cost drivers is better than an intercept-only model.

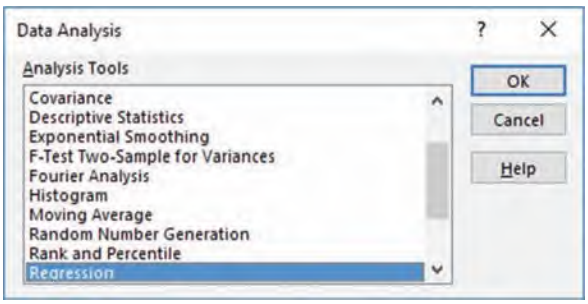
In fact, there are two cost drivers, miles per delivery and time per delivery, that seem to explain overhead costs. Notice that the *p*-value is less than 5% (alpha) for both of these cost drivers, suggesting we have found two cost drivers that impact overhead costs per delivery. We also note that weight per delivery does not have a *p*-value less than 5% (alpha), suggesting that it does not impact overhead costs per delivery.

Analysis Task 3: Apply the Relevant Cost Drivers to Future Deliveries.

We will now develop a formula to apply to future deliveries. We rerun the regression with the two cost drivers that have been shown to impact total overhead cost: miles per delivery and time per delivery.

1. Select **Data > Analyze > Data Analysis > Regression** (Lab Exhibit 3.5.7).

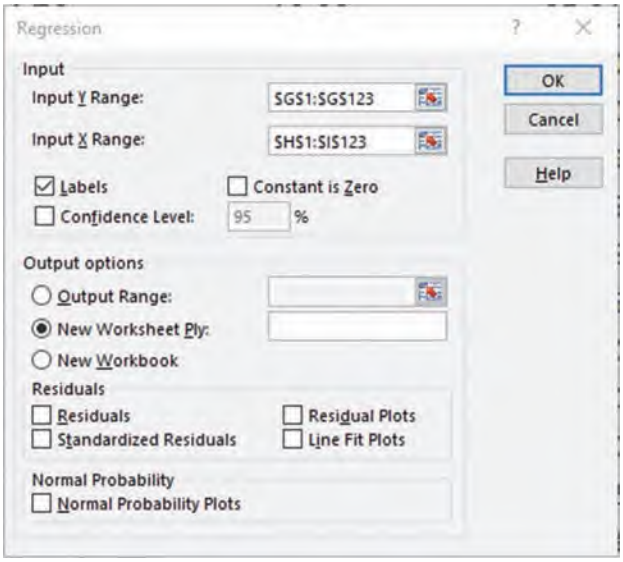
Lab Exhibit 3.5.7



2. In the regression dialog box, select the dependent variable (or Y Range) as overhead cost per delivery. Enter "$G$1:$G$123" in the Input Y Range.
3. Select the independent variable (or X Range). Our cost drivers (miles per delivery, time per delivery) are our independent variables, so enter $H$1:$I$123 in the Input X Range.
4. Because we highlighted the labels in the data range, click on "Labels". Also select New Worksheet Ply to get the results in a new worksheet.

The completed Regression dialog box should appear as shown in Lab Exhibit 3.5.8.

Lab Exhibit 3.5.8



5. Click OK to generate the regression results shown in Lab Exhibit 3.5.9.

Lab Exhibit 3.5.9

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.896213					
5	R Square	0.803198					
6	Adjusted R Square	0.79989					
7	Standard Error	4.487075					
8	Observations	122					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*gnificance F*	
12	Regression	2	9778.38	4889.19	242.8344	9.88E-43	
13	Residual	119	2395.927	20.13384			
14	Total	121	12174.31				
15							
16		*Coefficients*	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	14.45387	2.533622	5.704825	8.67E-08	9.437045	19.47069
18	Miles per Delivery	0.461936	0.096311	4.796316	4.73E-06	0.271232	0.652641
19	Time per Delivery	1.892175	0.116541	16.23608	5.8E-32	1.661412	2.122939

With these regression results, we are ready to make our formula to estimate overhead costs based on our two cost drivers. The intercept gives us the base cost for each delivery of $14.454. Because the coefficient on miles per delivery is 0.462, the cost per mile is $0.462. And because the coefficient on time per delivery is 1.892, the cost per minute is $1.892.

So, if we know how many miles the delivery took, and how many minutes it took to make the delivery, we can estimate and apply the overhead costs using this formula:

Estimated Overhead Costs = $14.454 + ($0.462 * Miles) + ($1.892 * Minutes)

To see how well our estimate does at explaining the overhead costs, we compare actual overhead costs to our estimated overhead costs using a scatterplot.

6. In cell K1 (on the Data sheet), type the title "Estimated". In cell L1, type the title "Actual".
7. Insert the formula "=14.454+(0.462*H2)+(1.892*I2)" into cell K2, as shown in Lab Exhibit 3.5.10. The formula should automatically copy down the column. Verify that all of the numbers are formatted with two decimals. This formula uses the line equation with the identified cost drivers, miles and time, to calculate estimated overhead costs.

Lab Exhibit 3.5.10

K
Estimated
=14.454 +(0.462*H2)+(1.892*I2)



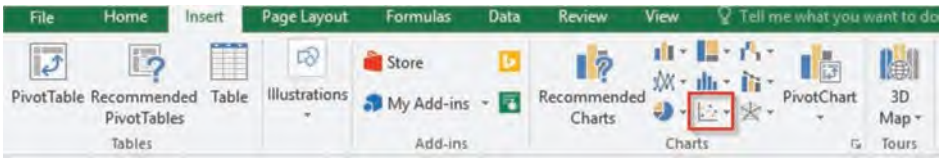
DATA VISUALIZATION

Report the Results: Visualize the Actual and Estimated Overhead Costs

To help present the results to management, it is often helpful to visualize the findings of the business analytics. Let's create a scatterplot comparing the estimated overhead costs from the regression (in column K) to the actual overhead costs.

1. Copy the values in G2:G123 to L2:L123, which puts the estimated next to the actual overhead costs per delivery. Placing these columns side by side facilitates the graphing of the two variables.

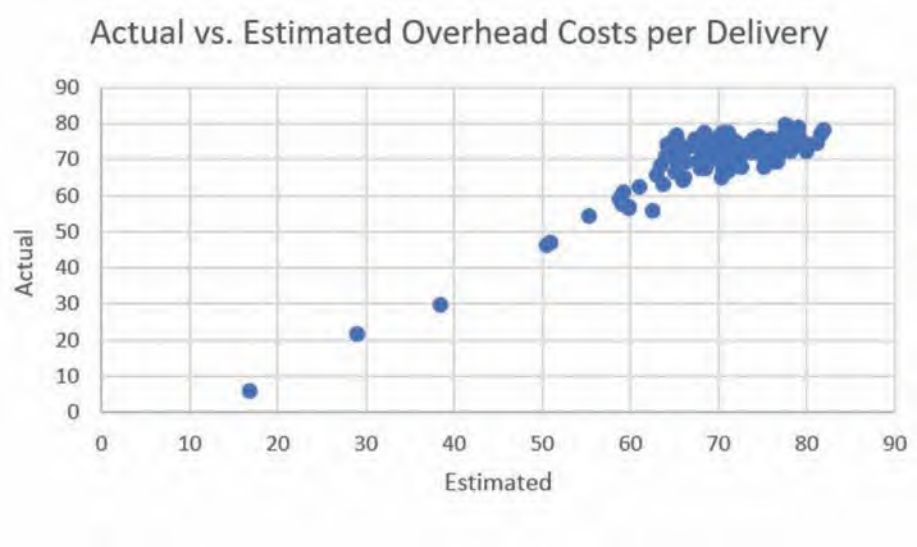
2. To graph the actual and estimated overhead costs, highlight cells K1:L123 and select **Insert > Charts > Scatter** or **Bubble Chart > Scatter** (Lab Exhibit 3.5.11).



Lab Exhibit 3.5.11

3. Change the title of the scatter plot to "Actual vs. Estimated Overhead Costs per Delivery".
4. Remove the zeros from the numbers in the X-axis and the Y-axis. For each axis, double-click on the axis. In the format Axis dialog box, expand Number, and set Decimal places to 0.
5. Add titles for the X-axis and Y-axis. Select **Chart Tools > Add Chart Element > Axis Titles > Primary Horizontal** (for X-axis) or **Chart Tools > Add Chart Element > Axis Titles > Primary Vertical** (for Y-axis). Label the X-axis "Estimated" and label the Y-axis "Actual". Lab Exhibit 3.5.12 shows the scatterplot.

Lab Exhibit 3.5.12



The result suggests that at least for this data set, there seems to be a pretty good correlation between actual and estimated overhead costs per delivery.

Assessment

1. Take a screenshot of your first regression output, copy it into a Word document named "Lab 3.5 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of your second regression output, copy it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the scatterplot, copy it into the same Word document, and label the screenshot Submission 3.
2. Answer the questions in Connect and upload your Lab 3.5 Excel Submission.docx via Connect, if assigned.

Alternate Lab 3.5 Excel

Deriving Cost Drivers for Activity-Based Costing (Regression Analysis)

Apply the same steps in Lab 3.5 to Lab 3.5 Alt Data.xlsx to perform cost driver analysis for the service call center at Thrustmaster. Thrustmaster is a fictitious company that sells game controllers, joysticks, steering wheels for PCs, and XBOX, Nintendo, and PlayStation consoles.

Thrustmaster is trying to understand its overhead allocation for the call center that supports its sales. It believes that by understanding the overhead costs and their respective cost drivers, the company will be in a better position to allocate overhead to the various products that require more service. Management is considering these four potential cost drivers:

1. Customer call minutes
2. Customer complaints
3. Average workers
4. Company sales

Required

1. Using regression analysis and choosing from several potential cost drivers, derive the cost drivers that are significant predictors of overall overhead costs and can be used to help allocate overhead. Set alpha to 5%, or 0.05.
2. Derive the formula for allocating overhead (the regression line equation).
3. Compare estimated overhead per delivery to actual overhead per delivery using a scatterplot.

Specify the Question

How can we use regression analysis to find cost drivers and derive a formula for allocating overhead?

Obtain the Data

This lab uses the data in Alt Lab 3.5 Data.xlsx. This data set contains information for potential cost drivers. The data dictionary for this file is below.

Data Dictionary

Customer Service Months:	The number of the month for the overhead costs
Total Overhead Cost:	The total overhead cost for the service month
Customer Call Minutes:	The number of minutes spent on customer calls during the month
Customer Complaints:	The number of customer complaints filed during the month
Average Workers:	The number of average workers during the month
Company Sales:	The gross sales revenue (in dollars) during the month

Assessment

1. Take a screenshot of your first regression output, copy it into a Word document named "Alt Lab 3.5 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of your second regression output, copy it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the scatterplot, copy it into the same Word document, and label the screenshot Submission 3.
2. Answer the questions in Connect and upload your Alt Lab 3.5 Excel Submission.docx via Connect, if assigned.

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Chapter 4

Analyze the Data:

Exploratory Business Analytics (Descriptive and Diagnostic Analytics)

A Look Back

- Chapter 3 introduced the statistical toolkit required to execute step 3 in the SOAR model: ANALYZE THE DATA.
- The toolkit includes sampling methods, descriptive (summary) statistics, and visualization methods and the use of statistical tests, hypothesis tests, correlations, and regression analysis to analyze the data.

A Look at This Chapter

- Chapter 4 introduces five types of business analytics, emphasizing the first two types of business analytics: descriptive analytics and diagnostic analytics as part of the third step of the SOAR model, ANALYZE THE DATA.
- Also called exploratory business analytics, descriptive and diagnostic analytics use specific analysis techniques, explained in this chapter.

A Look Ahead

- Chapter 5 emphasizes the use of predictive and prescriptive analytics in addressing business questions, explaining the third step of the SOAR model: ANALYZE THE DATA.
- Predictive and prescriptive analytics techniques include what-if analysis, cash flow analysis, marginal analysis, goal-seek analysis, and sensitivity analysis.



iStock / Getty Images

Understanding Life Expectancy

Suppose you work for **Fidelity** selling financial products (such as estate planning, insurance, and investment advice) to an older demographic. It is estimated that nearly 45 million U.S. households will transfer $68 trillion in wealth from baby boomers to the millennials over the next 25 years. You'd like to know the size and demographics of the 65-and-older population, both now and in the future, and what affects the size of that more senior population. Well, you're in luck!

The World Health Organization (WHO) has created a data set that includes life expectancy, adult mortality, and other demographic information for countries around the world and over time. Using this data set, you can use descriptive and diagnostic analytics to address your marketing team's many questions, exploring the relationships among numerous variables to uncover unknown and unexpected patterns as you attempt to answer the team's question: What variable has the biggest impact on life expectancy?

Here are some other questions that the marketing team might ask:

1. Does life expectancy have a positive or negative correlation with eating habits, lifestyle, exercise, smoking, drinking alcohol, and other factors?
2. How does schooling affect a person's life expectancy?
3. Do densely populated countries tend to have a lower life expectancy?
4. How, if at all, does immunization affect life expectancy?
5. How much has the COVID-19 pandemic affected life expectancy around the world?

Sources: https://towardsdatascience.com/14-data-science-projects-to-do-during-your-14-day-quarantine-8bd60d1e55e1 and https://www.cnbc.com/2018/11/20/great-wealth-transfer-is-passing-from-baby-boomers-to-gen-x-millennials.html. Accessed 10/2021. Link to the data set: https://www.kaggle.com/kumarajarshi/life-expectancy-who.

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

- **LO 4.1** Explain where ANALYZE THE DATA fits in the SOAR analytics model.
- **LO 4.2** Describe the five types of business analytics: descriptive, diagnostic, predictive, prescriptive, and adaptive/autonomous analytics.
- **LO 4.3** Define *exploratory analytics* and *confirmatory analytics*.
- **LO 4.4** Define *descriptive analytics* and enumerate its various techniques.
- **LO 4.5** Define *diagnostic analytics* and describe its techniques.
- **LO 4.6** Describe and use the techniques in Excel's Data Analysis ToolPak.

LABS FOR CHAPTER 4

- **LAB 4.1** **Excel:** Evaluating Inventory Using Inventory Turnover, Waste, and Profit Margins
- **LAB 4.2** **Excel:** Using Conditional Formatting to Perform a Bank Reconciliation
- **LAB 4.3** **Excel:** Applying Benford's Law

LO 4.1

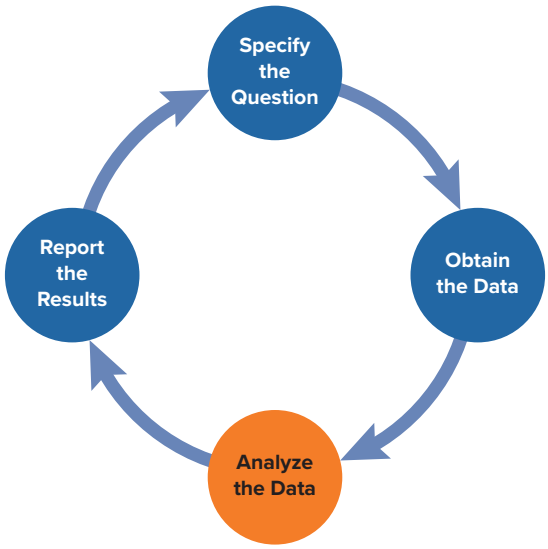
Explain where ANALYZE THE DATA fits in the SOAR analytics model.

4.1 THE THIRD STEP OF THE SOAR ANALYTICS MODEL: ANALYZE THE DATA

Recall from Chapter 1 that business analytics is the use of data to make knowledge, draw conclusions, and address business questions, which we operationalize through the SOAR analytics model (Exhibit 4.1). In earlier chapters, we discussed how to **S**PECIFY THE QUESTION (Chapter 1) and how to **O**BTAIN THE DATA (Chapter 2). Chapter 3 gave us the toolkit we need to analyze data. The next step in the SOAR analytics model is to **A**NALYZE THE DATA, which entails evaluating data to draw conclusions.

Analyzing data often includes manipulating the data and performing statistical analysis to answer the questions posed. This chapter, which is the first of two chapters focused specifically on analyzing the data, covers exploratory business analytics (descriptive and diagnostic analytics). Chapter 5 addresses confirmatory business analytics (predictive and prescriptive analytics). The final step of the SOAR model is **R**EPORT THE RESULTS, which we cover in Chapter 6.

Exhibit 4.1
The SOAR Analytics Model



PROGRESS CHECK

1. Does an analyst "Analyze the data" before or after finding the appropriate data to address a specific question? Explain.
2. Suppose you work for **Disney+** and you want to know which three Disney+ shows were most popular in Indianapolis last month. What specific data would help you answer this question?

LO 4.2

Describe the five types of business analytics: descriptive, diagnostic, predictive, prescriptive, and adaptive/ autonomous analytics.

4.2 MATCHING THE ANALYTICS TYPE TO THE BUSINESS QUESTION ASKED

There are many different types of analyses, which naturally leads to the question: What type of analytics do we use for what type of question? Not surprisingly, the type of questions we are asking determines the type of analytics we will use.

As we saw in Chapter 1, we generally group business questions into five types of relatively simple questions that correspond to the five types of analytics, as follows.[1]

1. **Descriptive analytics:** What happened? What is happening? (Chapter 4)
2. **Diagnostic analytics:** Why did it happen? What are the causes of past results? Why are the results different than expectations? (Chapter 4)
3. **Predictive analytics:** Will it happen in the future? What is the probability something will happen? Can we forecast what will happen? (Chapter 5)
4. **Prescriptive analytics:** What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints? (Chapter 5)
5. **Adaptive/Autonomous analytics:** How can we continuously learn using artificial intelligence? Can we learn from past and current events with adaptive learning? (Chapter 11)

The types of questions asked often vary with the role of the person asking. For example, a potential investor considering the financial payoffs of buying corporate bonds will ask different types of questions than a CEO who runs a billion-dollar business. Marketing managers' questions may focus more on questions related to creating new revenue streams than the questions of supply chain analysts or the **controller** (the individual who is responsible for all accounting-related activities at a company). A bank will have different questions about a company's performance than will the regulators at the Securities and Exchange Commission (SEC).

Exhibit 4.2 provides examples of specific business questions both by analytics type (descriptive, diagnostic, predictive, prescriptive, and adaptive/autonomous) and by the four main business functions (marketing, operations, accounting, and finance).

Business Function/ Analytics and Question Types	Marketing	Operations	Accounting	Finance
Descriptive Analytics • What happened? • What is happening?	• How many units were sold in the period? By region? By store? By country? • What was the average store revenue for each month of last year? • What are the top 10 products in terms of sales and profit? • How many webpage views resulted in a sale?	• What is this month's customer fill rate (percentage of customer orders we're able to fill without running out of product) for customers in Illinois? • What is the average supplier lead time? • What is the average customer service level? • What is the annual employee turnover rate? Which division has the highest turnover rate? • What is the average time-to-fill (an employee position)?	• How much did the company pay in federal taxes last year? • Which product is the most profitable one for the company? • What is the value of inventory on hand? • What is the average tax rate paid last year? • How big is the difference between taxable income and net income?	• What are the descriptive performance statistics (for example, means, medians, maximums, and minimums) for the retail industry? • What were the return on assets, return on profits, and asset turnover ratios last year?

[1]N. Tschakert, J. Kokina, S. Koziewski, and M Vasarhelyi. (2016). "The Next Frontier in Business Analytics," *Journal of Accountancy*, 222(2), 58.

(*continues*)

Business Function/ Analytics and Question Types	Marketing	Operations	Accounting	Finance
Diagnostic Analytics • Why did it happen? • What are the causes of past results? • Why are the results different than expectations?	• Why did revenue change? • Why did webpage views change? • Why did the click-through rate change from last month to this month?	• Why did customer fill rate change? • Why did average supplier lead time change? • Why did average customer service level change? • Why has employee turnover changed? • Why has time-to-fill changed?	• Why is revenue up on the West Coast but down in the Midwest? • Why did SG&A (sales, general, and administrative) expenses increase as compared to the industry? • Why is the difference between taxable income and net income getting bigger?	• Why did the interest rate increase for the company overall? • Why did stock returns increase last year in the United States as compared to China? • Why did bond prices fall?
Predictive Analytics • Will it happen in the future? • What is the probability something will happen? • Can we forecast what will happen?	• What will product sales (and units) be? • What will webpage views be next period? • Does the weather predict sales of certain products over other products?	• Can we forecast demand for our new product? • What happens if the company changes its inventory policy (for example, order quantity)? • How long will an employee stay with the company? • What will be the demand for our products based on new products from our competitors?	• Can we forecast future revenues/cash flows/ earnings with reasonable accuracy? • Can we predict if the financial statements might be misstated due to error or fraud?	• What is the chance the company will go bankrupt? • Do we extend credit or not to customers based on customer characteristics (such as credit score, payment history, and existing debt)?
Prescriptive Analytics • What should we do, based on what we expect will happen? • How do we optimize our performance based on potential constraints?	• What product recommendations might increase conversion rates? • How can we improve customer sentiment?	• What is the optimal inventory policy to maintain a specified item fill rate? What is the optimal level of production? • What is the optimal scheduling policy? • What is the best delivery route (in terms of costs and time)? • Where should warehouses and plants be located? • How many employees are needed for the summer season? • Which employees need to go to training to improve productivity (before productivity declines)?	• Should the company buy or lease its headquarters, office building? • Should the company make its products or outsource to other producers? • What is the level of sales that will allow us to break even? • Should the company move its operations to Ireland to minimize taxes or stay in the United States?	• Should the company move operations to Ireland to minimize taxes? • How will our taxes change if certain tax laws change in the next U.S. Congress legislative section? • What is the optimal investment: real estate, stocks, or bonds?

Business Function/ Analytics and Question Types	Marketing	Operations	Accounting	Finance
Adaptive/ Autonomous Analytics • How can we continuously learn using artificial intelligence? • How can we learn from past and current events with adaptive learning?	• Is our current marketing campaign working? How should we adjust our strategy today? • How should we adjust our strategy given the way we expect the competition/industry to change?	• How do we continuously update our predictions of needed inventory to more closely meet customer demand?	• Could a computer learn from the decisions of the internal audit department which transactions might constitute fraud (and require further investigation) and which ones will not?	• Can the company rebalance its investment portfolio to optimize its portfolio based on the change in oil prices last week? • What opportunities do changes in the exchange rates for foreign currencies present?

Exhibit 4.2
Business Analytics Type and Different Types of Questions by Four Business Functions (Marketing, Operational, Accounting, and Financial)

In this chapter, we focus on descriptive and diagnostic analytics, which we collectively call *exploratory business analytics.* In Chapter 5, we focus on predictive and prescriptive analytics, which we call *confirmatory business analytics.*

PROGRESS CHECK

3. Are diagnostic analytics (figuring out why something happened) harder to perform than descriptive analytics (figuring out what happened)? What makes these two types of analytics different from each other? What makes them similar?
4. Why is investigating why product recommendations affect conversion rates at **Walmart.com** (getting the people who read product reviews to actually buy the products reviewed) an example of diagnostic analytics instead of predictive analytics?

4.3 DEFINING EXPLORATORY AND CONFIRMATORY BUSINESS ANALYTICS

LO 4.3

Define *exploratory analytics* and *confirmatory analytics.*

It is useful to classify the types of business analytics according to their goal. **Exploratory business analytics** are the initial descriptive and diagnostic business analytics that analysts use to summarize and explain performance. In other words, exploratory business analytics explores historic data and begins to generate questions that require further investigation and hypotheses that need testing. Exploratory business analytics often includes describing past performance, exploring summary performance statistics, identifying anomalies and outliers, checking assumptions, and detecting patterns.

Confirmatory business analytics are predictive and prescriptive analytics that use statistics to judge the likelihood (probability) of a future event or outcome occurring. Before confirmatory data analytics can be used, it is important to understand existing relationships between values in the data set. We cover confirmatory data analytics in Chapter 5.

PROGRESS CHECK

5. Why is answering the question of why something occurred consistent with exploratory business analytics, as opposed to confirmatory analytics? Is it more difficult to predict future sales if you don't have an understanding of past sales? Explain.
6. Why is finding anomalies and outliers consistent with exploratory business analytics?

LO 4.4

Define *descriptive analytics* and enumerate its various techniques.

4.4 DESCRIPTIVE ANALYTICS

As we've seen, descriptive analytics addresses the questions of "What happened?" and "What is happening?" More formally, **descriptive analytics** characterize, summarize, and organize features and properties of the data to facilitate an understanding of business results and the underlying data.

Systems that Provide Data for Descriptive Analytics

Four principal business information systems at a company contain much of the data needed for descriptive analytics:

1. Human resource management systems (HRMS)
2. Customer relationship management (CRM) systems
3. Supply chain management (SCM) systems
4. Financial reporting (accounting) systems (FRS)

Each of these information systems may stand alone, or it may be included as a component of a company's integrated enterprise systems (also called *enterprise resource planning*). Let's examine each of these systems in a bit more detail.

Human Resource Management Systems (HRMS)

A **human resource management system (HRMS)** is an information system for managing all interactions with past, current, and potential employees. **BambooHR** is a popular vendor of HRMS, particularly for small businesses. Data included in the HRMS may include the following:

- Employee recruiting data and leads
- Employee training (current and desired certifications)
- Employee payroll and compensation (including tax elections such as 401(k) contributions, number of dependents, and amount of state and federal withholding)
- Employee benefits (stock options, bonuses, health insurance)
- Employee annual reviews (prior and current year reviews, ratings received)
- Employee absenteeism (how often work is missed)
- Employee career progression (prior and current roles)
- Employee satisfaction and sentiment survey results



Maintaining the privacy of employee data is critically important to a company. In addition, establishing and maintaining ethical standards regarding employee data collection and the use of employee data is a good way of establishing trust between employees and management.

Customer Relationship Management (CRM) Systems

A **customer relationship management (CRM) system** is an information system for managing all interactions with past, current, and potential customers. A popular vendor of CRM systems is **Salesforce**. CRM data may include the following:

- Customer contact history (What did the customer ask about or what was decided?)
- Customer order history (What did the customer order and when?)
- Customer level of trade discounts or payment terms (Which purchase incentives do we offer to customers?)
- Customer credit score (Is the customer's credit improving or declining?)
- Customer credit limit (How much credit is offered to the customer?)
- Customer payment history (How well or quickly are customers paying their bills?)

Like the privacy of employee data, the privacy of customer data is critically important to a company. Customers need to feel confident that their credit card details, credit scores, credit limits, and trade discounts remain confidential. For this reason, establishing and maintaining ethical standards regarding customer data collection, use, and storage is essential to maintaining good customer relations.



Supply Chain Management (SCM) Systems

A company's **supply chain** is the sequence of processes to move a product from raw materials to final customer delivery. The system used to track these processes is called a **supply chain management (SCM) system**. For example, **Tesla** may use supply chain systems to get a view of its products' movement along the supply chain from the factory to the final customer. **SAP** is a popular vendor of supply chain systems. A supply chain system may include data related to:

- Active vendors (their contact information, where payment should be made, how much should be paid)
- The orders placed to date (how much is made, when the orders are made)
- Demand schedules for which components of the final product are needed, along with when and where
- Transportation status and schedules
- The current location of the product in the supply chain

Financial Reporting Systems (FRS)

Financial reporting systems (FRS), or accounting systems, capture and measure financial transactions and communicate financial performance to interested parties. **Oracle** is a popular vendor of financial reporting systems. The output of FRS typically includes the **financial statements**, which are a collection of reports that communicate a company's financial results, financial condition, financial health, and cash flows.

Financial statements, such as the **balance sheet** (which summarizes a company's assets and how those assets are financed at a certain point in time) and **income statement** (which summarizes how profitable the company has been over a period of time), are a systematic means of reporting performance. They allow accountants and financial advisers to specifically address the question "What happened?" Additional external sources of financial data may also be available to help address financial performance questions, including macroeconomic statistics gathered and published by the government, industry-level reports, and competitor analysis (including analysis of their financial statements).

Statistical and Summarization Techniques for Descriptive Analytics

Summarization and statistical techniques to address the basic questions "What happened?" or "What is happening?" are generally all found in spreadsheet software (such as Microsoft Excel) and statistics software (such as SAS and SPSS). Exhibit 4.3 provides brief descriptions and examples of these descriptive analytics techniques, which include counts, totals, sums, averages, subtotals, minimums, maximums, histograms, pivot tables, standard deviations, and ratio analytics. These techniques characterize, summarize, and organize the available data.

Descriptive Analytics Techniques	Use	Example Questions
Counts	Show how frequently an event or process occurs	How many units were sold in the period, by region and by store, last year?
Totals, sums, averages, subtotals, and cross-tabulations (such as pivot tables)	Summarize measures of performance	What is the average supplier lead time? What is the average customer service level?
Minimums, maximums, medians, standard deviations	Summarize measures showing extreme values to help explain what happened	What are the descriptive performance statistics (such as means, medians, maximums and minimums) for the retail industry's sales last year?
Graphs (bar charts), histograms, and other visualizations	Summarize visually how often or how much there is of an object or event in comparison to other values	What is the average time to fill an employee position over time? How has it changed over time?
Vertical analysis and horizontal analysis of financial statements	Summarize changes in revenues from one period to the next	How much did cost of goods sold as a percent of revenues increase over the past two years?
Ratio analyses such as return on assets, return on sales (profit margin), asset turnover ratios, or debt-to-equity ratios	Calculate important financial ratios to evaluate a company's performance	What are the descriptive performance ratios (for example, return on assets, return on sales, asset turnover ratios, debt-to-equity ratios) for **Walmart**, **Target**, and **Amazon** last year?

Exhibit 4.3
Statistical Techniques Used in Descriptive Analytics

Using Descriptive Statistics as a Descriptive Analytics Technique

Descriptive statistics are summary statistics that briefly describe or summarize features of a data set or variable. Descriptive statistics include the average, standard deviation, minimum, and maximum.

Lab 4.1 Excel, titled "Evaluating Inventory Using Inventory Turnover, Waste, and Profit Margins," uses descriptive analytics such as mean (average), median, maximum, minimum, and other statistics, as shown in Exhibit 4.4 for the Exotic Fruits company.

LAB CONNECTION

LAB 4.1 Excel asks you to work with descriptive statistics and descriptive analytics, specifically considering the inventory turnover, waste, and profit margins for a set of products.

	A	B	C	D
1	*Inventory Turnover*		*Waste as a % of Costs*	
2				
3	Mean	5.6611	Mean	0.1361
4	Standard Error	0.4554	Standard Error	0.0123
5	Median	4.4673	Median	0.1167
6	Mode	#N/A	Mode	#N/A
7	Standard Deviation	4.0983	Standard Deviation	0.1105
8	Sample Variance	16.7964	Sample Variance	0.0122
9	Kurtosis	3.4888	Kurtosis	2.9268
10	Skewness	1.7758	Skewness	1.5115
11	Range	20.7960	Range	0.5707
12	Minimum	1.2805	Minimum	0.0021
13	Maximum	22.0765	Maximum	0.5728
14	Sum	458.5487	Sum	11.0250
15	Count	81	Count	81

Exhibit 4.4
Descriptive Statistics for Exotic Fruits (From Lab 4.1 Excel)

Data Visualization in Descriptive Analytics: Graphs and Histograms


DATA VISUALIZATION

Data visualizations are often used to illustrate and understand past performance. For example, as mentioned in Chapter 1, line graphs can show the trend of customer sales over the most recent eight quarters (or any other period of time). A pie chart can show the composition of those customer sales (such as international sales or wholesale versus retail sales). A bar chart can compare total sales by month of the year to help managers understand seasonality.

Histograms are also used to show a distribution of values. A **histogram** presents rectangles whose area is proportional to the underlying frequency of the data. Exhibit 4.5 is a histogram that reports gross margin percentages for Exotic Fruits by showing the relative frequency of products sold in each of four gross profit margin percentage ranges.

Gross margin percentage: Net sales minus the cost of goods sold all divided by net sales. Gross margin percentage is the percent of money a company retains after deducting the direct costs of producing a good or service.

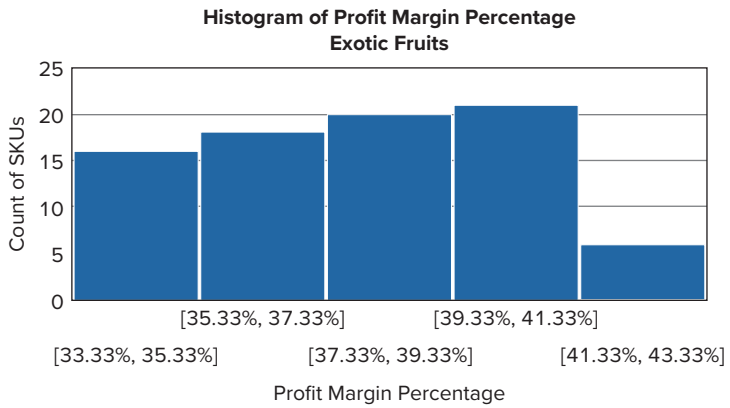


Exhibit 4.5
Histogram of Gross Profit Margin Percentages at Exotic Fruits (by Number of Products Sold in Each Gross Profit Margin Percentage Range)

Examples of Descriptive Analytics: Horizontal and Vertical Analysis of Performance

Two common types of descriptive analyses of performance are horizontal analysis and vertical analysis. Although our examples in the following sections use data from an income statement, these analyses can be performed with almost any type of data that look at a change in the performance of an entity or unit over time.

Horizontal Analysis

Horizontal analysis identifies changes in various line items of the income statement (for example, cost of goods sold or interest expense), balance sheet (for example, inventory or long-term notes payable), or statement of cash flows (for example, change in accounts receivable or sale of common stock) across time. In other words, horizontal analysis calculates and displays the dollar and percentage change in financial statement line items from one period to the next. It is called "horizontal" analysis because it looks at changes across a row in a spreadsheet, doing so horizontally from one year to the next. Horizontal analysis is also called *trend analysis* because it helps identify trends across time.

Exhibit 4.6 illustrates a horizontal analysis for **Tyson Foods Inc**. performed using Excel. Cell D6 calculates the change in dollar sales by taking the most current year's sales (in this case, 2020) and subtracting the sales in the previous year (2019). This same calculation is performed for all items in the income statement as shown in column D.

In cell E6, to calculate the percentage change in sales, we take the most current sales (2020), subtract the sales in the previous year (2019) and divide by the sales in the earlier year. We can express the formula as:

$$\text{Percentage change (increase or decrease)} = \frac{(\text{New} - \text{Old}) * 100}{\text{Old}}$$

We then perform this same calculation in column E for all income-statement line items.

Horizontal analysis helps us see the specific changes from one year to the next, allowing us to summarize clearly and in detail what in fact happened.

	A	B	C	D	E
1	Tyson Foods Inc (TSN)				Percent
2	Consolidated Statements Of Income ($ millions)			Change from	Change from
3	*Fiscal Period*	**2020**	**2019**	**2020-2019**	**2020-2019**
4	Sales	$47,049	$43,185	$3,864	8.95%
5	Cost of Sales	$40,523	$37,801	$2,722	7.20%
6	Gross Profit	$6,526	$5,384	$1,142	21.21%
7	**Operating Expenses:**				
8	Selling, General and Administrative	$2,130	$2,376	($246)	-10.35%
9	Operating Income	$4,396	$3,008	$1,388	46.14%
10	**Other (Income) Expense:**				
11	Interest income	($8)	($10)	$2	-20.00%
12	Interest expense	$428	$485	($57)	-11.75%
13	Other, net	($65)	($131)	$66	-50.38%
14	Total Other (Income) Expense	$355	$344	$11	3.20%
15	Income before Income Taxes	$4,041	$2,664	$1,377	51.69%
16	Income Tax Expense (Benefit)	$981	$593	$388	65.43%
17	Net Income	$3,060	$2,071	$989	47.75%

Exhibit 4.6
Horizontal Analysis for Tyson Foods Inc. Income Statement

Source: Tyson Foods 2021 10-K, Accessed December 2021.

LAB CONNECTION

Lab 8.1 Excel (in Chapter 8) provides an illustration of horizontal analysis for **Apple** and **Google**. It also demonstrates the use of sparklines to quickly display what occurred. A sparkline is a data visualization that fits within one cell in Excel and visually depicts what happened in the cells next to it.

Vertical Analysis

Another method of descriptive analytics that we can use to examine financial performance over a period of time is **vertical analysis**, which expresses financial information in relation to some relevant figure, or base. For example, on an income statement, the relevant base is always net sales (or net revenue). On a balance sheet, the relevant base is always total assets. It is called vertical analysis because each item is a proportion of another number (such as sales or total assets), and all of the items are listed in the same column vertically.

Exhibit 4.7 is a vertical analysis of **Tyson Foods**' (Ticker Symbol = TSN) income statement. For each year, each line item (such as Cost of Sales) is expressed as a percentage of net sales for that year. For 2021, we divide the Cost of Sales of $40,523 by total 2021 sales of $47,049 to get 86.13%. For 2020, the result of the calculation is 87.53%, so it appears that Cost of Sales as a percentage of net sales is decreasing.

Also note that Selling, General and Administrative expenses decreased from 5.50% of revenues in 2020 to 4.53% of revenues in 2021. This decrease of almost 1% could be considered a good trend, suggesting that Tyson is doing a better job controlling these expenses.

Tyson Foods, Inc. (TSN)				
Consolidated Statements Of Income			Percent of Sales	
	2021	2020	2021	2020
Sales	$47,049	$43,185	100.00%	100.00%
Cost of Sales	$40,523	$37,801	86.13%	87.53%
Gross Profit	$6,526	$5,384	13.87%	12.47%
Operating Expenses:				
Selling, General and Administrative	$2,130	$2,376	4.53%	5.50%
Operating Income	$4,396	$3,008	9.34%	6.97%
Other (Income) Expense:				
Interest income	($8)	($10)	-0.02%	-0.02%
Interest expense	$428	$485	0.91%	1.12%
Other, net	($65)	($131)	-0.14%	-0.30%
Total Other (Income) Expense	$355	$344	0.75%	0.80%
Income before Income Taxes	$4,041	$2,664	8.59%	6.17%
Income Tax Expense	$981	$593	2.09%	1.37%
Net Income	$3,060	$2,071	6.50%	4.80%

Exhibit 4.7
Vertical Analysis for Tyson Foods Inc. Income Statement

Source: Tyson Foods 2021 10-K, Accessed March 2022.

LAB CONNECTION

Lab 8.2 Excel (in Chapter 8) provides an illustration of vertical analysis for **Apple** and **Google**. It also demonstrates the use of sparklines as a visualization to quickly display what occurred. A sparkline is a data visualization that fits within one cell in Excel and visually depicts what happened in the cells next to it.

PROGRESS CHECK

7. Why are vertical analysis and horizontal analysis both considered to be descriptive analytics techniques?
8. What visualization type would you use to show a trend over time? Why is this considered to be descriptive analytics?

LO 4.5

Define *diagnostic analytics* and describe its techniques.

4.5 GOING FROM DESCRIPTIVE ANALYTICS TO DIAGNOSTIC ANALYTICS

Sometimes analysts who perform descriptive analytics may find an outcome that is quite different from what they expected. Their initial intent had been to find out what happened, but the process triggers the need for diagnostic analytics to explain that outcome. This is one reason why descriptive analytics are often performed before diagnostic analytics, even though there is no set process requiring analytics to be performed in a specific order or to include all types of analytics.

While descriptive analytics addresses the question "What happened?" diagnostic analytics takes the analysis a step further by asking "Why did it happen?" and "What are the causes of past results?" **Diagnostic analytics** are performed to investigate the underlying reasons for past results that cannot be answered by simply looking at the descriptive data.

Diagnostic analytics usually fall into two broad categories[2]:

1. **Identifying anomalies and outliers**. Often the first step in diagnostic analytics is to look for and identify unusual, unexpected results or transactions. An **anomaly** is something that departs, or deviates, from the expected or the norm. An **outlier** is an observation, or a data point, that lies outside its expected distribution. Anomalies and outliers could be mistakes, truth, or the result of fraud. After identifying anomalies and outliers and evaluating what might have happened, analysts investigate further to understand why they occurred.
2. **Finding previously unknown linkages, patterns, or relationships between variables**.
   a. **Performing drill-down detailed analytics**. By looking for patterns in the underlying data set or an external data set, business analysts can identify potential correlations. Drill-down analytics works to summarize data at different levels and uncover additional details to help analysts understand why something happened.
   b. **Performing statistical analyses**. This category of business analytics works to uncover statistical patterns in the data, or how data move together.

Diagnostic Analytics: Identifying Anomalies and Outliers

The first broad category in diagnostic analytics is identifying anomalies and outliers. Based on the results of descriptive analytics, business analysts may investigate the results that do not make sense compared to the expected results.

Management by Exception

The common management practice of **management by exception**, which entails managers' attention to specific aspects of the business that depart from expectations, is consistent with identifying anomalies and outliers. For example, as long as marketing expenses are

[2]Diagnostic Analytics 101: Why Did It Happen? IBM, https://www.ibm.com/blogs/business-analytics/diagnostic-analytics-101-why-did-it-happen/. Accessed February 14, 2019.

within budget and meeting expectations, there is little reason for managers to investigate them any further. However, when marketing expenses are over budget or not achieving expected results, managers may address these specific exceptions directly, with the goal of understanding why the results differed from their expectations.

Establishing Expectations to Help Identify Anomalies and Outliers

To detect anomalies and outliers in diagnostic analytics, it is important to establish the norm, or the expectation. For example, the marketing department might expect the company's sales to increase at roughly the same rate as sales in other companies in the same industry. If sales do not increase in line with the competition, then this anomaly should be investigated further.

Exhibit 4.8 provides examples of what the business analyst (or management) might expect, along with examples of possible anomalies and possible explanations.

Business Function	The Business Analyst's Expectation	Example of Anomaly or Outlier	Possible Explanation for the Anomaly/Outlier
Marketing	The increase in sales is roughly equivalent to the increase in sales for other companies in the same industry.	A company's sales grew at twice the rate of its closest competitor.	The new products sold particularly well, causing an increase in sales greater than the competition's.
Accounting	Accountants record transactions in conformity with Generally Accepted Accounting Principles (GAAP).	An auditing company found that its client capitalized rather than expensed research and development (R&D) expenditures, contrary to GAAP.	The company bought long-term testing equipment for future research and development. Because this equipment is expected to last more than one year, the client argues it should be capitalized for future projects.
Finance	A company's bond prices are priced just lower than U.S. Treasury bonds.	A company's bond prices fell significantly in relation to U.S. Treasury bonds.	Due to a company's poor financial performance, its bonds were graded as risky, thus lowering the bond price.
Operations	A company's employee turnover rate this year will be roughly equivalent to employee turnover rate last year.	Employee turnover is 25 percent higher than it was last year.	Other businesses in the industry need skilled workers and are offering better compensation packages.

Exhibit 4.8
Anomalies and Outliers: Expectations, Examples, and Explanations

An Illustration of Diagnostic Analytics Techniques for Identifying Anomalies and Outliers

As noted earlier, an anomaly may occur due to error or fraud or for some other justifiable or non-justifiable reason. Certain statistical techniques are used as part of diagnostic analytics to look for and identify unusual, unexpected results or transactions. These techniques include the following:

1. Looking for duplicates or unique values using Excel's MATCH() function or **conditional formatting** (application of specific formatting to cells that meet certain criteria), as in a cash/bank reconciliation (see Lab 4.2 Excel).
2. Using data visualizations, such as graphs, charts, scatterplots, and data sorts, to identify minimum and maximum values of a set of numbers, and to assist in finding extreme or unexpected values.

Exhibit 4.9 provides additional examples of different diagnostic analytics techniques used to identify potential anomalies. Let's look more closely at specific techniques.

Diagnostic Analytics Technique	Expectation	Potential Anomaly (or Departure from Expectation)	How Technique Works to Identify Anomaly
Cash/Bank reconciliation (reconciles a company's bank account with its financial records)	All cash transactions recorded by the company are also recorded by the bank.	Some cash transactions are recorded by the company or by the bank but not both, due to checks that aren't yet cleared by the bank (outstanding checks, etc.).	Excel's conditional formatting and MATCH() function highlight transactions not recorded by both (see Lab 4.2 Excel). Bank reconciliations are used to make sure all transactions are recorded properly.
Benford's Law	According to Benford's Law, the first digit of naturally occurring numerical data sets follow an expected distribution.	The leading digit of some naturally occurring data set does not follow the expected distribution, suggesting something that requires further investigation.	The Benford's Law technique compares first digits of numbers in the data set to first digits of numbers expected by Benford's Law (see Lab 4.3 Excel). Benford's Law is often used to identify fraud and/or errors.
Testing for duplicate transactions	Each transaction is recorded once and only once.	Some transactions are duplicated, leading to errors.	A duplicate transaction test looks for the same amount recorded two or more times in order to correct possible errors.
Fuzzy matching (a technique used to find potential equivalents when there is less than an exact fit).	Different names represent different individuals. Different addresses represent different locations.	Although names or addresses might not be exactly the same, they actually represent the same person.	Fuzzy matching attempts to find a match that, although not a 100 percent match, is above the threshold matching percentage set by the analyst (see Lab 8.3 Excel).
Sequence check (a technique to determine if a key number field in the record is in the correct ascending or descending order)	The identifying number for each transaction has a different identifying number, and these numbers follow a prespecified order.	Some identifying numbers and the transactions they represent are skipped, suggesting errors.	A sequence check looks for missing transactions, which are then investigated as to the reason for the skipped transaction.

Exhibit 4.9
Different Diagnostic Analytics Techniques Used to Identify Potential Anomalies

An Example of Diagnostic Analytics: Bank Reconciliations

One common diagnostic analytics technique for identifying anomalies is reconciling the transactions affecting the cash account in the company's **general ledger** (the primary accounting records for a company) to the bank statement. The *bank reconciliation* identifies transactions on the bank statement that do not correspond with the company's general ledger and vice versa.

Potential reconciling items include:

- *Outstanding checks:* Checks written by the company but not yet processed by the bank.
- *Outstanding deposits:* Deposits recorded by the company but not yet processed by the bank.

- *Non-sufficient funds (NSF) checks:* Checks deposited by the company but not paid by the bank because the check writer has insufficient funds.
- *Electronic funds transfer* (*EFT, wire transfer*): Transfers received by the bank but not yet reflected in the company's records.
- *Notes and interest:* Notes are loans made to bank customers and collected by the bank. *Interest* is the cost of borrowing, usually expressed as a percentage of the loan, and payable by the borrower. The company may not be aware of the amounts of these items until it receives the bank statement.
- *Bank service fees:* The bank charges fees for its services—for example, a monthly fee for each checking account. The bank deducts these fees directly from the company's checking account. The company may not account for the fee until it receives the monthly bank statement.
- *Errors:* Sometimes a company or its bank makes a mistake in recording a transaction. Analysts find these errors by comparing the amounts for each transaction on the general ledger and each transaction on the bank statement.

Let's now look at a specific anomaly:

EXPECTATION	ANOMALY
According to Benford's Law, the first digit of naturally occurring numerical data sets follows an expected distribution.	The leading digit of some naturally occurring data set does not follow the expected distribution, suggesting something that requires further investigation.

Why is this an example of an anomaly? Any items not included in both the general ledger and the bank statement are potentially fraudulent or erroneous, or they might need further analysis to be fully explained. Exhibit 4.10 provides an example of the use of conditional formatting to perform a bank reconciliation. The items highlighted in blue match, but the unhighlighted items need to be reconciled. The items that need reconciliation should be investigated further. Note how the visualizations included as part of conditional formatting allows the analyst to quickly see where the exceptions are.



DATA VISUALIZATION

LAB CONNECTION

Lab 4.2 Excel uses diagnostic analytics to perform a bank reconciliation. Specifically, the lab uses conditional formatting to highlight which items need to be reconciled between the cash transactions (checks, deposits, and so on) recorded by the company in its general ledger and those recorded by the bank.

Another Example of Diagnostic Analytics: Benford's Law

Benford's Law is the principle that in any large, randomly produced set of natural numbers, there is an expected distribution of the first, or leading, digit. According to Benford's Law, a higher percentage of numbers in a population of numbers starts with 1 rather than any other digit, followed by those that begin with 2, then 3, and so on, as shown in Exhibit 4.11. Under Benford's Law, numbers starting with 1 should account for about 30 percent of leading digits, and each successive number should represent a progressively smaller proportion, with 9's coming last at just under 5 percent of the leading digits.[3]

[3]To get the expected distribution for each digit, we use the equation $\log_{10}(1 + (1/d))$, where d is each digit. So, the number 1 is expected to be the first digit 30.1 percent of the time, the number 2 is expected to be the first digit 17.61 percent of the time, and so on.

Exhibit 4.10
Example of Bank Reconciliation Performed Using Conditional Formatting (As Used in Lab 4.2 Excel)

	A	B	C	D	E
1	Family First Cabins				
2	Cash Transactions				
3	March 31, 2024				
4	General Ledger/ Company Books			Bank Statement	
5					
6	Transaction Type			Transaction Type	
7	Deposit 1023	2,073.14		Deposit 1024	107.98
8	Deposit 1024	107.98		Deposit 1028	2,985.10
9	Deposit 1025	289.75		Deposit 1027	1,819.13
10	Deposit 1026	699.46		Deposit 1025	289.75
11	Deposit 1027	1,819.13		Deposit 1026	699.46
12	Deposit 1028	2,985.10		Deposit 1023	2,073.14
13	Deposit 1029	993.09		Deposit 1029	993.09
14	Deposit 1030	3,010.23		Deposit 1030	3,010.23
15	Deposit 1031	1,064.20		Check 11762	1,414.13
16	Deposit 1032	1,065.38		Check 11764	2,041.73
17	Deposit 1033	2,613.29		Check 11765	987.34
18	Check 11762	1,414.13		Check 11766	1,501.19
19	Check 11763	11,783.98		Check 11767	2,278.18
20	Check 11764	2,041.73		Check 11768	4,398.01
21	Check 11765	987.34		Check 11769	1,349.56
22	Check 11766	1,521.19		Check 11770	2,430.61
23	Check 11767	2,278.18		Note Collect 1	1,500.00
24	Check 11768	4,398.01		Interest Earned	11.05
25	Check 11769	1,349.56		NSF Check 1013	2,645.05
26	Check 11770	2,430.61		NSF Check 1017	924.12
27	Check 11771	1,513.19		NSF Fees 1	25.00
28	Check 11772	2,985.09		NSF Fees 2	65.00
29	Check 11773	865.23			
30	Check 11774	152.90			

When a distribution of numbers departs from Benford's Law, that distribution is an anomaly that should be investigated further. For example, in Exhibit 4.12, the difference between the blue bars that represent the actual distribution of first digits in a data set and the expectation given Benford's Law (as shown by the orange line) is statistically tested. Using a specific statistical test, we measure how the expectations spelled out in Benford's Law compare to actual observed data. If the *p*-value of the chi-square test is very small depending on our value of α (alpha), say 0.01 or 0.05, then the observed relationship is different than Benford's Law would predict. (For a review of these statistical concepts, see Chapter 3.) This data visualization of the difference between actual and expected percentages may help the analyst find the source(s) of the difference. The difference may indicate fraud or an error that requires further investigation.



DATA VISUALIZATION

EXPECTATION	ANOMALY
The reconciled cash balance at the bank is the same amount as the reconciled cash balance in the general ledger.	A bank reconciliation (comparison of bank statement and general cash account) reveals items unaccounted for by the bank, the company, or both.

First Digit	Expected Probability Distribution of First Digits (%)
1	30.10
2	17.61
3	12.49
4	9.69
5	7.92
6	6.69
7	5.80
8	5.12
9	4.58

Exhibit 4.11
Benford's Law: Expected Probability Distribution by First Digit of Naturally Occurring Numbers

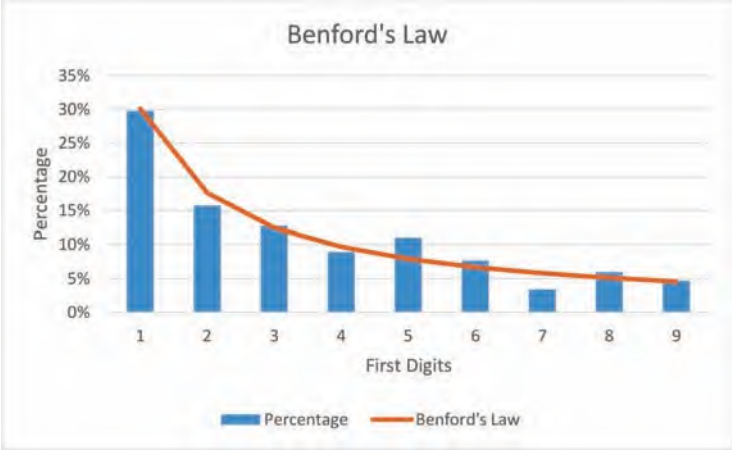


Exhibit 4.12
Comparing the Actual Occurrence of the First Digit of a Set of Numbers with the Predictions Gathered Using Benford's Law (from Lab 4.3 Excel)

LAB CONNECTION

LAB 4.3, Excel provides an example of the Benford's Law technique to determine if the actual observed data of first digits differ from the distribution predicted by Benford's Law (as shown in Exhibit 4.12).

MINI CASE: Benford's Law in Practice

Let's look at an example of Benford's Law in practice and how diagnostic analytics can help a company detect and correct a problem. Several hundred customer service representatives at a company's call center were authorized to issue refunds of up to $50 to customers, without additional approval. Any refund larger than $50 required a supervisor's approval. Because each customer service representative had processed more than 10,000 refunds over several years—most without any supervisor approval—there was the distinct chance that fraud had occurred. **KPMG**, a major accounting firm, worked with the call center's managers to check the validity of the refunds using Benford's Law.

In investigating the call center, KPMG retrieved the relevant data from the database administrator and evaluated the first digits of the refunds issued by each customer

service representative. It then calculated the frequencies and compared them with the expected distribution.

For certain customer service representatives, the refunds did not follow Benford's Law. In fact, for some representatives, leading digits of 4 were much more numerous than all other digits, which led KPMG to believe the operators might be giving out too many refunds just below the $50 threshold. Indeed, KPMG discovered that customer service representatives were giving out many refunds in the $40–$49.99 range.

The accountants took their analytics a step further and identified a handful of representatives who had issued fraudulent refunds to themselves, friends, and family that totaled several hundred thousand dollars.

Decision-Making with Data

How might data analysts use Benford's Law be to identify fraud in the following situations? (a) a public election, (b) tax returns, (c) stock market trades, (d) cryptocurrency transactions, (e) audits of financial statements.

Source: J. C. McGinty, "Accountants Increasingly Use Business Analytics to Catch Fraud," *Wall Street Journal,* 2014, https://www.wsj.com/articles/accountants-increasingly-use-data-analysis-to-catch-fraud-1417804886.

Finding Previously Unknown Linkages, Patterns, or Relationships Between Variables

We can also use diagnostic analytics to find the extent to which patterns exist between and among variables in a data set. Two common techniques are drill-down analytics and statistical analysis.

Drill-Down Analytics

In **drill-down analytics**, we uncover additional details in the data by digging more deeply into the data and summarizing the data at different levels to understand why something happened. In doing so, we look for patterns among variables to identify potential relationships and patterns. Using drill-down analytics, we shift from an overview of the data to a more detailed (or granular) view of the data to gain insight.

Gross margin: Net sales minus the cost of goods sold. Gross margin is the amount of money a company retains after deducting the direct costs of producing a good or service.

Think of drill-down analytics as successively peeling off the layers of an onion. After you've peeled the first layer of the onion, you see the next layer, evaluate it, and address it. You then keep drilling down through successive layers, analyzing and draw conclusions about them, until you get to the center of onion.

Exhibit 4.13 summarizes drill-down analytics techniques and provides examples of the types of questions addressed and specific techniques employed. These drill-down techniques look for patterns in the underlying data set or an external data set to identify potential correlations with **cross-tabulations** (pivot tables), data sorts, and data manipulations.

Drill-Down Analytics Technique	Example Question	Specific Technique Employed
Uncover the details by summarizing the data at different levels	Which employee was the most productive in each division and in each store?	Summarize employee productivity metric(s) by division and by store.
Use cross-tabulations (method of quantitatively analyzing the relationship between two or more variables) to view transactions from a different perspective	Which car model at **General Motors** has the highest gross margin?	Use cross-tabulations with a pivot table to evaluate which products are most profitable.

Exhibit 4.13
Techniques and Examples of Performing Drill-Down Analytics as Part of Diagnostic Analytics

MINI CASE: Drill-Down Techniques in Sports: The Oakland A's

As chronicled in the 2011 movie *Moneyball,* Billy Beane and Peter Brand of the Oakland A's used a sophisticated drill-down diagnostic analytics approach to scout and analyze Major League baseball players. (Some people watch the movie to learn about data analytics or baseball in general; others just like Brad Pitt.)

The analysts were not trying to find the best players overall, but rather the best players per dollar of salary. Their goal was to assemble a full team of strong players rather than overpaying for superstars. Who could throw the most strikeouts per dollar of salary, the most home runs per dollar of salary, and so on?

Similarly, other business can drill down into their data to find ways to attract their best and second-best customers in an economically feasible way, or to find and reward their most productive employees per dollar of pay. They can also evaluate the effectiveness of their advertising, such as including web ads, trade shows, and online shopping channels, per marketing dollar spent.

Decision-Making with Data

Suppose you want to identify the highest percentage NBA 3-point shooter per dollar of salary and the quarterback with the highest NFL quarterback rating per minute played. Which data sources would you use?

Statistical Analysis

Consider the statistical techniques, example questions, and specific techniques employed to determine relations/patterns/linkages between variables as shown in Exhibit 4.14. Some of the terms in the exhibit will not be familiar to you yet. You will learn more about them in later chapters.

Descriptive analytics is often performed before diagnostic, predictive, or prescriptive analytics. However, there is no set process requiring business analytics types be done in a specific order or to include all steps. Sometimes when the analyst is performing descriptive

Statistical Technique to Determine Relations/Patterns/Linkages Between Variables	Example Question	Specific Technique Employed
Correlation (the extent to which variables are related to each other)	To what extent are inventory problems at **Wayfair** associated with its number of supply-chain partners?	Run a correlation analysis to assess the relationship between inventory depletion and number of suppliers.
Regression (an assessment of how a specific outcome is related to specific inputs)	Which factors at **Accenture** (pay, training, convenient work hours, ability to work remotely) are associated with employee turnover?	Run a regression analysis to assess the relationship between factors that may be associated with employee absenteeism and employee turnover.
Hypothesis testing (a statistical test of an assumption or a theory)	Is the time to source semiconductors from Philippines shorter than the time to source the same product from Vietnam?	Use a statistical test to compare the average cycle time (how long it takes to complete a customer's order) of orders from the Philippines to average cycle times of orders from Vietnam.

Exhibit 4.14
Statistical Techniques, Example Questions, and Specific Techniques Employed to Determine Relations/Patterns/Linkages Between Variables

analytics, he or she finds an outcome, or result, that seems quite different from what was expected. So, while the intent of the analysis was to find out what happened (that is, descriptive analytics), sometimes the analysis triggers the need for diagnostic analytics to explain why it happened. For example, horizontal or vertical analysis, a descriptive analytics technique, might identify an issue, and then diagnostic analytics could be used to find an explanation as to why production increased but overall revenues did not (for example).

Although there is no prescribed order for performing the various analytics types, you may have a hard time forecasting future sales (predictive analytics) or optimizing decision-making for the future (prescriptive analytics) if you do not know and understand what has happened with past and current sales (descriptive and diagnostic analytics).

PROGRESS CHECK

9. You've developed a hypothesis that it is faster to source artificial Christmas trees from Indonesia (to Seattle) than from the Philippines (to Seattle). Using statistical analysis, how would you test that hypothesis? Why is this an example of diagnostic analytics?
10. Suppose that an examination of banking transactions at **Wells Fargo** reveals more check amounts starting with 9's than with 3's. Do these findings match what Benford's Law predicts? Why or why not?
11. **Southwest Airlines** has a frequent flier program called Rapid Rewards, which gives its customers free flights after accumulating points from flying on Southwest Airlines. Which type(s) of analytics would be needed to understand if the benefits of running the Rapid Rewards program are greater than the costs of administering the program and giving free flights?

LO 4.6

Describe and use the techniques in Excel's Data Analysis ToolPak.

4.6 INTRODUCTION TO TECHNIQUES USED IN DATA ANALYSIS

In this book, we emphasize the use of Excel and Tableau to perform data analysis to address business questions. Additional tools are available to help with all the components of the SOAR model except for the SPECIFY THE QUESTION phase. Excel can be used to extract and manipulate data, analyze data, or share the story by visualizing data. Tableau emphasizes data visualizations, recognizing that visualizations are often the best way to explore and present data.

Exhibit 4.15 summarizes the types of analytics and techniques/functions that you can perform in Excel. Some of the terms in this table may not be familiar to you. Don't worry about that for now; you will learn more about them in later chapters. To use some of the Excel functions, you will need to download Microsoft Excel's Data Analysis ToolPak.

Data Analytics and Your Career

Do you need to learn Excel for your career? The answer is YES. Thousands of career postings require proficiency in Excel. The top nine are:

1. Accountants and auditors
2. Administrative assistants
3. Business, management, and market analysts
4. Cost estimators

5. Educators
6. Financial analysts
7. Market research analysts
8. Project managers
9. Sales, marketing, training, and administrative managers

Source: "Excelling in Your Career: 9 Top Careers That Require Microsoft Excel Skills," https://www.alliancecareertraining.com/excelling-in-your-career-9-top-careers-that-require-microsoft-excel-skills/. Accessed September 21, 2021.

Business Analytics Type	Specific Types of Analysis Employed	Microsoft Excel Techniques/Functions
Descriptive	Descriptive statistics such as counts, totals, sums, averages Histograms, box plots, graphs, and charts	SUM(), COUNT(), COUNTIF(), AVERAGE(), SUBTOTAL(), MEDIAN() Pivot tables; quartiles, quintiles, deciles; percentiles Data Analysis ToolPak: histogram, descriptive statistics
Diagnostic	Variances, differences from expectations, correlations, identify outliers/anomalies (fuzzy lookup) Sequence testing, drill-downs, and roll-ups to get detail when needed Cross-tabulations Principal component analysis clustering Hypothesis testing	Pivot tables Data Analysis ToolPak: Statistical testing of means of various groups (*t*-test), rank and percentile Conditional formatting Excel add-in: Fuzzy lookup

Exhibit 4.15
Business Analytics Type, Specific Analysis Employed, and Excel Techniques/Functions

The Excel Data Analysis ToolPak

Excel offers a Data Analysis ToolPak that helps you perform many of the business analytics that we use in this text. If you haven't installed the Data Analysis ToolPak yet, Appendix E at the end of this text provides instructions for doing so.

If you aren't sure if you've installed the Excel Data Analysis ToolPak, look for it on the **Data > Analysis** ribbon (as shown in Exhibit 4.16 and Lab 1.3 Excel). The appearance may vary by Mac, PC, and Excel version.

To perform additional data analysis, select **Data > Analysis > Data Analysis**. A dialog box will open, as shown in Exhibit 4.17.

In this text, we use the following aspects of Excel's Data Analysis ToolPak (for a refresher, see Chapter 3):

Correlation: To understand if and the extent to which variables are related to each other
Descriptive statistics: To understand the basic statistics, including the mean (average), standard deviation, and minimums and maximums, of a data set

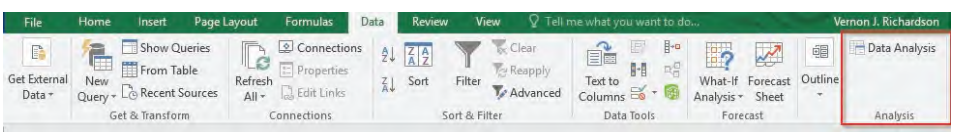
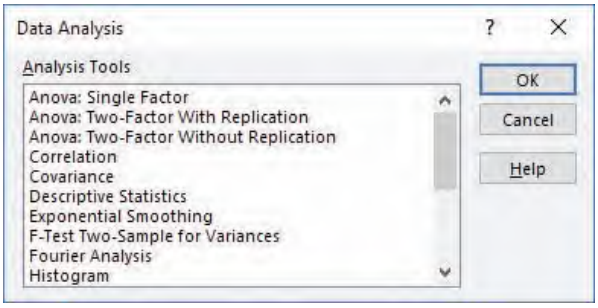


Exhibit 4.16
Finding the Data Analysis ToolPak in Excel

Exhibit 4.17
Analytics Techniques Available Within the Excel Data Analysis ToolPak



Histogram: To understand the frequency of the data using a display of rectangles whose areas are proportional to the underlying frequency of the data
Regression: To understand the relationship between specific inputs and outputs
***t*-test:** To understand the probability of a statistical difference in averages between two independent samples or between a paired sample through time

PROGRESS CHECK

12. Which Excel Data Analysis ToolPak technique would be most useful in forecasting the level of sales a firm will experience, given its level of advertising expense? Would this analysis be considered predictive analytics or descriptive analytics?
13. Give two scenarios where you might employ the Excel Data Analysis ToolPak's descriptive statistics. To help identify the scenarios, consider some of the questions posed in Exhibit 4.2.

Chapter Summary

The type of analytics that analysts perform often depends on the type of question asked. We discussed five types of questions and five corresponding analytics types: descriptive, diagnostic, predictive, prescriptive, and adaptive/autonomous analytics. We also discussed the analytics techniques employed to address the various questions asked.

In this chapter, we highlighted the use of exploratory business analytics, emphasizing descriptive and diagnostic analytics and the related techniques. Descriptive analytics use summary statistics and other statistics to provide a summary of what happened. Diagnostic analytics addresses the question "Why did it happen?" by finding anomalies and outliers, and by using other drill-down and statistical techniques to uncover unknown patterns and linkages among variables.

We concluded by looking at Excel's Data Analysis ToolPak, which we will use in the analytics techniques we perform throughout the rest of the book.

Key Terms

anomaly (*232*) Something that departs, or deviates, from the expected or the norm.

balance sheet (*227*) A financial statement that summarizes a company's assets and how those assets are financed at a certain point in time.

bank reconciliation (*234*) A diagnostic analytics technique that evaluates differences between a company's bank account and its financial records.

Benford's Law (*235*) The principle that in any large, randomly produced set of natural numbers, there is an expected distribution of the first, or leading, digit. A higher percentage of numbers in a population of numbers starts with 1 rather than any other digit, followed by those that begin with 2, then 3, and so on.

conditional formatting (*233*) Application of specific formatting to cells (in Excel or Google Sheets) that meet certain criteria.

confirmatory business analytics (*225*) Predictive and prescriptive analytics that use statistics to judge the likelihood (probability) of a future event or outcome occurring.

controller (*223*) The individual who is responsible for all accounting-related activities at a company.

correlation (*239*) Business analytics technique to understand if, and the extent to which, variables are related to each other.

cross-tabulations (*238*) A descriptive analytics technique to quantitatively analyze the relationship between two or more variables.

customer relationship management (CRM) system (*227*) An information system for managing all interactions with past, current, and potential customers.

descriptive analytics (*226*) Analytics that characterize, summarize, and organize features and properties of the data to facilitate an understanding of business results and the underlying data.

descriptive statistics (*228*) A descriptive analytics technique that uses summary statistics that briefly describe or summarize features of a data set or variable, including the average, standard deviation, minimum, and maximum.

diagnostic analytics (*232*) Analytics performed to investigate the underlying reasons for past results that cannot be answered by simply looking at the descriptive data.

drill-down analytics (*238*) A descriptive analytics technique that evaluates a greater level of detail in the data to gain insight.

exploratory business analytics (*225*) Initial descriptive and diagnostic business analytics that analysts use to summarize and explain performance.

financial reporting systems (*227*) Systems that capture and measure financial transactions and communicate financial performance to interested parties; also called accounting systems.

financial statements (*227*) A collection of reports that communicate a company's financial results, financial condition, financial health, and cash flows.

fuzzy matching (*234*) A diagnostic analytics technique used to find potential equivalents when there is less than an exact fit.

general ledger (*234*) The primary accounting records for a company that summarize the impact of all the company's transactions (sales, expenses, investments, and so on).

histogram (*229*) Data visualization that uses rectangles whose area is proportional to the underlying frequency of the data.

horizontal analysis (*230*) A descriptive analytics technique that calculates and displays the dollar and percentage change in financial statement items across time (from one period to the next period); also called trend analysis.

human resource management system (HRMS) (*226*) An information system for managing all interactions with past, current, and potential employees.

hypothesis testing (*239*) A statistical test of an assumption or theory.

income statement (*227*) A financial statement that summarizes how profitable a company has been over a period of time.

management by exception (*232*) A management technique in which managers attend to specific aspects of the business that depart from expectations.

outlier (*232*) An observation or data point that lies outside its expected distribution.

regression (*239*) Technique used to understand the relationship between specific inputs and outputs.

sequence check (*234*) A diagnostic analytics technique to determine if a key number field in the record is in the correct ascending or descending order.

supply chain (*227*) The sequence of processes that moves a product from raw materials to final customer delivery.

supply chain management (SCM) system (*227*) The system that coordinates the flow of goods and services from original production to final customer delivery.

vertical analysis (*231*) A descriptive analytics technique that expresses financial information in relation to some relevant figure, or base (total sales for the income statement, and total assets for balance sheet items) from the same year.

ANSWERS TO PROGRESS CHECKS

1. ANALYZE THE DATA is the third step of the SOAR analytics model. Once we know our question (SPECIFY THE QUESTION) and acquire and clean the appropriate data (OBTAIN THE DATA), we are ready to perform the appropriate analytics to address the question.
2. To determine the most popular show on Disney+ in Indianapolis, you must first OBTAIN THE DATA and then ANALYZE THE DATA. Data summarizing the streaming of shows according to zip code or metropolitan area would help you rank Disney+ shows in order of popularity.
3. Descriptive and diagnostic analytics answer different types of questions and use different techniques. Diagnostic analytics is not necessarily harder to perform than descriptive analytics; however, diagnostic analytics requires us to ask more questions than descriptive analytics and query additional data sources to determine why something did or did not happen in comparison with expectations.
4. Diagnostic analytics entails answering the question "Why did it happen?" Trying to explain why some people bought the products after reading reviews at Walmart.com, and other people did not, is a good example of diagnostic analytics.
5. Often it is important to explore the details of a relationship before that relationship can be confirmed. For example, it will be much more difficult to predict future sales (confirmatory analytics) if you cannot see prior sales trends (exploratory analytics).
6. Finding and analyzing data to identify anomalies and outliers is an important way to understand how and why a business has performed the way it has (exploratory business analytics).
7. Horizontal analysis and vertical analysis both analyze what occurred, addressing the question "What happened?"—which is consistent with descriptive analytics.
8. A line graph can be used to show a trend over time. Evaluating trends over time addresses the question "What happened?", which is consistent with descriptive analytics.
9. Hypothesis testing is one technique used in diagnostic analytics. It allows us to test a theory or our "gut instinct" to determine whether a phenomenon is statistically significant. In diagnostic analytics, we seek to understand the root cause or determinant of a phenomenon. A *t*-test could be used to test the hypothesis, comparing the sourcing time from Indonesia to Seattle to the sourcing time of Philippines to Seattle.

10. According to Benford's Law, in any large, randomly produced set of natural numbers, there is an expected distribution of the first, or leading, digit. A higher percentage of numbers in a population of numbers starts with 1 than any other digit, followed by those that begin with 2, then 3, and so on. Thus, unless there is a specific explanation for why there are more leading digits with 9's instead of 3's, we'd want to investigate the situation to explain why this phenomenon is occurring.
11. The benefits of the frequent flier program likely include loyal customers who choose to fly on Southwest Airlines rather than other carriers, thereby increasing Southwest's revenue. Diagnostic analytics may be used to figure out how much additional revenue Southwest takes in as a result of its Rapid Rewards program. Descriptive analytics could assess the costs associated with administering the program and giving away free flights.
12. Correlation and regression can be used to determine the relationship between advertising expenses and the sales they generate. Both of these techniques are available in the Excel Data Analysis ToolPak. Forecasting sales given the level of advertising expenses is an example of predictive analytics. However, we generally need to perform diagnostic analytics to uncover the relationships before we forecast the future as part of predictive analytics.
13. Descriptive statistics are generally consistent with descriptive analytics explaining what happened or diagnostic analytics looking at extreme values to find anomalies and outliers. Descriptive statistics could be used to answer questions such as "What is this month's average employee turnover rate for our stores in Illinois?" and "What was the average store revenue for each month of last year?"



Multiple-Choice Questions

1. (LO 4.2) Which question does predictive analytics address?
   a. What happened?
   b. What should we do based on what we expect will happen?
   c. Why did it happen?
   d. Will it happen in the future?
2. (LO 4.5) Which type of question does diagnostic analytics address?
   a. What happened?
   b. What should we do based on what we expect will happen?
   c. Why did it happen?
   d. Will it happen in the future?
3. (LO 4.2) What type of analytics addresses the question "What happened?"
   a. Descriptive analytics
   b. Diagnostic analytics
   c. Predictive analytics
   d. Prescriptive analytics
4. (LO 4.4) To compute the percentage change in a horizontal analysis from last year to this year, the formula is:
   a. (Current Year – Last Year) * 100 / Last Year
   b. (Last Year – Current Year) / Last Year
   c. (Last Year – Current Year) / Current Year
   d. (Current Year – Last Year)/ Current Year

5. (LO 4.4) In vertical analysis, what is the relevant base for cost of goods sold?
   a. Net sales revenue
   b. Total assets
   c. Total stockholders' equity
   d. Total credit sales
6. (LO 4.4) Which of the following expresses financial information in relation to some relevant figure or base?
   a. Vertical analysis
   b. Horizontal analysis
   c. Ratio analysis
   d. A pivot table
7. (LO 4.4) Descriptive analytics
   a. summarizes the relevant data.
   b. uncovers patterns in the relevant data.
   c. conducts what-if scenarios regarding the relevant data.
   d. performs cash-flow analysis for the relevant data.
8. (LO 4.4) From which system would a company likely get data identifying the customer with the highest dollar amount of orders in the past month?
   a. Customer relationship management system
   b. Accounting system
   c. Human resource management system
   d. Supply chain management system
9. (LO 4.5) According to Benford's Law, which one of these first digits has a lower probability of occurrence than the first digit "5"?
   a. 3
   b. 2
   c. 4
   d. 9
10. (LO 4.5) Evaluating an uncertain relationship between advertising expenditures and net income each year is an example of
    a. performing drill-down detailed analytics.
    b. determining relations/patterns/linkages between variables through statistical analysis.
    c. identifying anomalies/outliers.
    d. performing descriptive analytics.
11. (LO 4.3) Which type of analytics is more likely to summarize and explain prior performance.
    a. Exploratory
    b. Probabilistic
    c. Confirmatory
    d. Advanced
12. (LO 4.6) Conditional formatting is most often used as a technique in which type of analytics?
    a. Diagnostic
    b. Descriptive
    c. Predictive
    d. Prescriptive

Discussion Questions

1. (LO 4.2) Why does each question that a decision-maker might ask require a different type of analytics (descriptive, diagnostic, predictive, prescriptive, or adaptive/autonomous analytics)?
2. (LO 4.2) Describe the differences between predictive analytics and adaptive/autonomous analytics.
3. (LO 4.4) Why is understanding what happened (descriptive analytics) important to understanding why something happened (diagnostic analytics)?
4. (LO 4.2) Which type of analytics answers the question "How much in state taxes did the company pay during the previous year?"
5. (LO 4.2) Which type of analytics emphasizes totals, sums, averages, and sum totals? Provide an example.
6. (LO 4.2) Which type of analytics works to uncover previously unknown relationships and linkages among variables? Provide an example.
7. (LO 4.3) Contrast exploratory and confirmatory business analytics. How are they different from each other?
8. (LO 4.4, 4.5) Suppose we are interested in investigating whether a social media advertising campaign was more effective than a traditional media campaign in creating additional revenues. To answer this question, which should we use: descriptive analytics or diagnostic analytics? Support your answer.
9. (LO 4.5) How does a bank reconciliation use anomalies and outliers to find items that need to be reconciled? Which type of analytics does a bank reconciliation represent?
10. (LO 4.6) Given the role of descriptive analytics in addressing the questions "What happened?" and "What is Happening?" how do the Excel commands SUM(), COUNTIF(), and AVERAGE() help analysts answer those questions?



Brief Exercises

1. (LO 4.2) Match the analytics type (descriptive, diagnostic, predictive, prescriptive, adaptive/autonomous) to the question it addresses.

Question	Analytics Type
How can we improve customer satisfaction?	
What is the expected product demand over the next five years?	
Why did our competition perform better than we did?	
Based on expected demand for our products, how should we price our products to optimize profits?	
Is our current marketing campaign working?	
Should we move our operations to Malaysia to minimize taxes ,or should our operations remain in the United States?	

2. (LO 4.4, 4.5) Match each analytics type (descriptive or diagnostic) to the techniques used.

Analytics Technique Used	Descriptive Analytics or Diagnostic Analytics?
Counts	
Correlation	
Sum total	
Horizontal analysis	
Regression	
Standard deviation	
Vertical analysis	
Bank reconciliations	

3. (LO 4.4, 4.5) Match each analytics type (descriptive or diagnostic) to the techniques used.

Technique/Analysis Used	Descriptive Analytics or Diagnostic Analytics?
Ratio analysis	
Benford's Law	
Percentage change	
Pivot table to find the most profitable product or customer	
Sequence check	
Fuzzy matching of vendors and employees to see if there are similarities in names or addresses	
Testing for duplicate transactions	

4. (LO 4.6) Match the business analytics technique in Excel's Data Analysis ToolPak to its use.

Business Analytics Technique	Use
1. Correlation	a. To understand the probability of a statistical difference in averages between two independent samples or between a paired sample through time
2. Regression	b. To understand if and the extent to which variables are related to each other
3. Descriptive statistics	c. To understand the basic statistics, including the average, standard deviation, and minimums and maximums of a data set
4. Histogram	d. To understand the frequency of the data using a display of rectangles whose areas are proportional to the underlying frequency of the data
5. *t*-test	e. To understand the relationship between specific inputs and outputs

Problems

1. (LO 4.5) Name the diagnostic test analytics technique used to find each type of anomaly. Choose from these techniques: bank reconciliations, Benford's Law, fuzzy matching, sequence check, and testing for duplicate transactions.

Anomaly	Diagnostic Test Approach/Technique
Cash transactions recorded by the company are different from those recorded by the bank.	
We suspect that certain sales were recorded twice	
There are more transactions with the leading digit of 8 in the data set than those with the leading digit of 3.	
An invoice number was skipped.	
The address of an employee is similar to the address of a vendor.	

2. (LO 4.5) Name the diagnostic analytics technique used to answer the following business questions. Choose from this list: correlation, regression, hypothesis testing.

Business Questions	Diagnostic Analytics Technique (Correlation, Regression, or Hypothesis Testing)
What is the relationship between product cost and product quality?	
Is the cycle time of sourcing vanilla from Madagascar shorter than the cycle time of sourcing vanilla from Tahiti?	
Do female executives receive lower levels of compensation than male executives?	
What factors are associated with employees' calling out sick?	
What factors are associated with *Fortune* 500 companies' stock returns?	

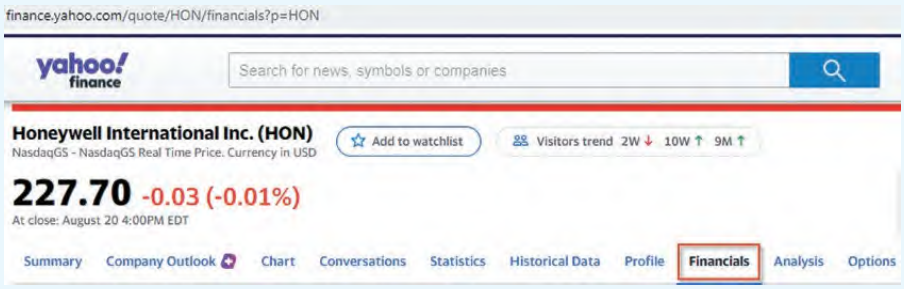
3. (LO 4.2) Describe specific types of questions asked in marketing for a specific company of your choice that might require descriptive, diagnostic, predictive, prescriptive, and adaptive/autonomous analytics, with at least one question for each. (Use Exhibit 4.2 as a guide.) Describe specific techniques that could be used to address each of those questions.
4. (LO 4.4) Using **Alphabet Inc. (Google)** 2020 and 2021 results (as shown below), perform a horizontal analysis to calculate the change in results from 2020 to 2021. *Note:* All numbers are in millions.

Required

a. Calculate the change in revenues and the percentage change in revenues from 2020 to 2021.
b. Calculate the change in net income and the percentage change in net income from 2020 to 2021.
c. Compare the percentage change in revenues and percentage change in net income. Did you expect them to be similar? Do your calculations support your expectations? Explain.

	A	B	C
1	Alphabet Inc. (GOOG)		
2	CONSOLIDATED STATEMENTS OF INCOME		
3	*Fiscal Period*	2021	2020
4	Revenues	$257,637	$182,527
5	Costs and expenses:		
6	Cost of revenues	$110,939	$84,732
7	Research and development	$31,562	$27,573
8	Sales and marketing	$22,912	$17,946
9	General and administrative	$13,510	$11,052
10	Total costs and expenses	$178,923	$141,303
11	Income from operations	$78,714	$41,224
12	Other income (expense), net	$12,020	$6,858
13	Income before income taxes	$90,734	$48,082
14	Provision for income taxes	$14,701	$7,813
15	Net income	$76,033	$40,269

5. (LO 4.4) Go to Yahoo! Finance for **Honeywell** (HON) (https://finance.yahoo.com/quote/HON/financials?p=HON). Then go to the Financials tab (as shown below) to access the income statement. Perform vertical analysis on Honeywell's income statement for the two most recent years. What insights can be found?



finance.yahoo.com/quote/HON/financials?p=HON

6. (LO 4.4) According to a recent press release (September 2021), "Statistical analyses used to detect fraud in elections held in Third World nations (such as Russia, Venezuela, and Iran) have detected fraud in California resulting in Governor Gavin Newsom being reinstated as governor. The primary analytical technique used was Benford's Law and [the results] can be readily reproduced."[4] How would you design an experiment using Benford's Law to identify fraudulent elections? What numbers can you evaluate to assess first digits? Voter ID? Votes in each polling station?
7. (LO 4.6) Download the data set, Sales Growth Dataset.xlsx, which has sales data (in $millions) for years 20X1 and 20X2 for a set of publicly traded companies over a recent year.

Required

1. Format the two Sales columns to two digits.
2. Calculate the sales growth using the formula [(New – Old) * 100] / Old or the formula [($Sales_{20X2}$ – $Sales_{20X1}$) * 100] / $Sales_{20X2}$. Format the Sales Growth Column to three digits.
3. Run descriptive statistics using the Data Analysis ToolPak in Excel. How many observations are in the data set? Also, identify the maximum, minimum, median, average, and range of the sales growth data. Round your answer to three decimals.

[4]Source: https://abc7.com/larry-elder-voter-fraud-california-recall-gavin-newsom/11019011/.

8. (LO 4.4) Download the Unpaid Invoices data set in Excel from Connect. Unpaid invoices refers to the money owed to the company by its debtors (typically its customers). In this problem, we are going to describe the unpaid invoices using different descriptive analytics techniques, including quartiles and percentiles.

Required

1. Using the Quartile function in Excel, calculate the minimum and maximum values and the value of the first quartile (25th percentile), second quartile (50th percentile), and third quartile (75th percentile).
2. Using the Percentile function in Excel, calculate the value of the 1st, 17th, 47th, 64th, and 99th percentile.
3. Is there a difference in values between the 50th percentile, the second quartile, and the median? Explain.
4. How many observations are in the data set?
5. What tool/technique within Excel could you use to accumulate the unpaid invoices by customer?

9. (LO 4.5) Which has the highest return on investment—the S&P 500, high-grade corporate bonds, or U.S. Treasury bonds? Download the Investment Returns data set in Excel. The source is Aswath Damodaran, NYU Finance Professor, at http://pages.stern.nyu.edu/~ adamodar/New_Home_Page/datafile/histretSP.html.

Required

1. Calculate the average returns for the S&P 500 stock index, high-grade corporate bonds, and U.S. Treasury bonds from 1928 to 2020. Which has the highest return?
2. Calculate the standard deviation of returns to the S&P 500 stock index, high-grade corporate bonds, and U.S. Treasury bonds from 1928 to 2020. Which has the highest standard deviation of returns? (*Hint:* Use the STDEV Excel function for each index.
3. Generally we expect that the higher the risk, the greater the expected return. Which of these three investments has the highest risk (measured here as standard deviation of returns), and which has the greatest return?

10. (LO 4.3, 4.4) Which has the highest return on investment from 1928 to 2020—the S&P 500, high-grade corporate bonds, or U.S. Treasury bonds? Download the Investment Returns data set in Excel. The source is Aswath Damodaran, NYU Finance Professor, at http://pages.stern.nyu.edu/~ adamodar/New_Home_Page/datafile/histretSP.html.

Required

1. Using the Analysis ToolPak in Excel, perform a *t*-test: Two-Sample Assuming Equal Variances between the annual returns on the S&P 500 stock index and high-grade corporate bonds. (*Hint:* Set Hypothesized Mean Difference to 0 (zero) and alpha at 0.05). What is the *p*-value associated with the *t*-test? Is it less than our established alpha of 0.05? Is one statistically greater than the other (using a one-tailed test)?
2. Using the Analysis ToolPak in Excel, perform a *t*-test: Two-Sample Assuming Equal Variances between the annual returns on high-grade corporate bonds and U.S. Treasury Bonds. (*Hint:* Set Hypothesized Mean Difference to 0 (zero) and alpha at 0.05.) What is the *p*-value associated with the *t*-test? Is it less than our established alpha of 0.05? Is one statistically greater than the other (using a one-tailed test)?
3. What can you conclude about the returns to these different investments?
4. Given the results above, what do you expect would be the result of a *t*-test between the S&P 500 stock index and U.S. Treasury bonds?



DATA VISUALIZATION

Data Visualization Exercises

1. (LO 4.5) Benford's Law: To establish the expectation for the frequency of leading digit of a naturally occurring data set, we use the equation $\log_{10}(1 + (1/d))$, where d is each digit 1, 2, and so on.

Required

1. Calculate the expected distribution for each digit using Benford's Law.
2. Plot the results on a graph to visualize the expected distribution for all of the leading digits in a distribution. Which digit has the lowest expected percentage of observations?

2. (LO 4.4) Download the data set, Sales Growth Dataset.xlsx, which has sales data for year 20X1 and 20X2 for a set of publicly traded companies. Calculate the sales growth using the formula (New – Old) / Old or the formula $(Sales_{20X2} - Sales_{20X1})/Sales_{20X2}$ for each company.

Required

1. Use Excel, Tableau, or Power BI to plot the results in a column chart, histogram, or box plot. Which of these data visualization types will be most useful in diagnostic analytics to find the most extreme sales growth percentages? Explain.
2. Does the visualization suggest that the sales growth percentages skew to the left or to the right? What do you conclude from the results?

3. (LO 4.4) Use the following data to graph the sales revenue and net income for **Walmart** from 2012 to 2020 in Excel. The data file, Walmart Trend Analysis.xlsx, is available in Connect or via the Additional Student Resources page. (Note: The data are in $ millions.)

Year	Sales Revenue	Net Income
2012	$468,651	$17,756
2013	$476,294	$16,695
2014	$485,651	$17,099
2015	$482,130	$15,080
2016	$485,873	$14,293
2017	$500,343	$10,523
2018	$514,343	$ 6,670
2019	$523,964	$14,881
2020	$559,151	$13,510

Required

1. Create a bar graph for sales revenue, and create a line graph for net income, formatting the X- and Y-axes with year and numbers and inserting appropriate chart titles.
2. What trends do you see for Walmart sales and Walmart net income over this time period?

LAB 4.1 **Excel:** Evaluating Inventory Using Inventory Turnover, Waste, and Profit Margins
LAB 4.2 **Excel:** Using Conditional Formatting to Perform a Bank Reconciliation
LAB 4.3 **Excel:** Applying Benford's Law

Lab 4.1 Excel

Evaluating Inventory Using Inventory Turnover, Waste, and Profit Margins

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Ratio Analysis, Inventory Turnover

Decision-Making Context

Determining how much inventory to hold in stock is an important business decision. In this lab, we use the **inventory turnover ratio** and the extent of waste to assess the proper inventory levels.

Inventory turnover ratio: The number of times inventory is sold in a year.

Exotic Fruits, a fictitious company, stocks a wide inventory of fruits from around the world. The company is hoping to lower its costs and enhance the profitability of the fruits it chooses to carry.

Required

1. Compute the inventory turnover ratio.
2. Compute waste as a percent of cost.
3. Create a visualization showing waste as a percent of cost.

Specify the Question

How much inventory, and which inventory items, should Exotic Fruits hold?

Obtain the Data

Exotic Fruits has been keeping detailed data on its inventory, including sales price and cost per pound, average inventory, cost of goods sold, and waste/loss.

1. Open Excel file Lab 4.1 Data.xlsx. The data dictionary appears below.

Data Dictionary

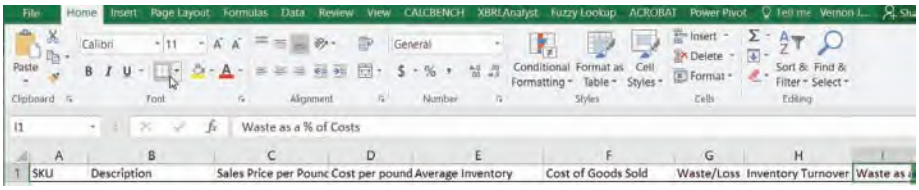
SKU:	Exotic Fruits' SKU (An SKU, or stock-keeping unit, is a unique code for each fruit sold)
Description:	Description of each fruit sold
Sales Price per Pound:	Total sales price per pound of product
Cost per Pound:	Total cost of product sold per pound of product
Average Inventory:	The average inventory on hand during the year (in pounds)
Cost of Goods Sold:	The total cost of each fruit over the year
Waste/Loss:	The dollar loss of fruit wasted

Analyze the Data

Analysis Task 1: Compute the Inventory Turnover Ratio and Waste as a Percent of Costs.

1. We will compute the inventory turnover ratio to determine which inventory turns over fastest. To compute the inventory turnover ratio, we divide the cost of goods sold by average inventory.
2. In cell H1, type "Inventory Turnover". In cell I1, type "Waste as a % of Costs". Select cells H1 and I1, and insert a border underneath both cells, as shown in Lab Exhibit 4.1.1.

Lab Exhibit 4.1.1



3. In cell H2, insert the formula "=F2/E2" and select Enter, as shown in Lab Exhibit 4.1.2.

Lab Exhibit 4.1.2

	A	B	C	D	E	F	G	H	I
1	SKU	Description	Sales Price per Pounc	Cost per pound	Average Inventory	Cost of Goods Sold	Waste/Loss	Inventory Turnover	Waste as a
2	FRT-APL	Apple	1.7	1.09	51.067	908.792	42.91	=F2/E2	
3	FRT-APO	Apricot	2.8	1.76	30.631	298.030	10.15		
4	FRT-AVD	Avocado	1	0.59	20.054	288.724	5.74		

Copy the result from cell H2 to cells H3:H82. Highlight column H, and then right-click and format cells with "Number" and three decimal places.

4. In cell I2, insert the formula "=G2/F2" and select Enter, as shown in Lab Exhibit 4.1.3.

Lab Exhibit 4.1.3

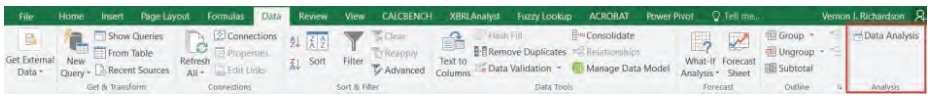
	A	B	C	D	E	F	G	H	I
1	SKU	Description	Sales Price per Pou	Cost per poun	Average Inventory	Cost of Goods Sold	Waste/Loss	Inventory Turnove	Waste as a
2	FRT-JAU	Jambul	1.95	1.13	7.600	16.238	1.7	2.137	=G2/F2
3	FRT-PUE	Purple mangosteen	1.01	0.59	7.530	22.256	5.35	2.956	
4	FRT-JAB	Jabuticaba	0.89	0.52	7.347	16.049	2.62	2.184	
5	FRT-PEA	Pear	1.83	1.07	35.954	270.328	15.69	7.519	
6	FRT-BIT	Bilberty	2	1.17	6.856	46.849	3.46	6.833	
7	FRT-MII	Miracle fruit	2.05	1.2	6.787	26.463	4.66	3.899	
8	FRT-JAI	Jackfruit	0.63	0.37	5.018	6.515	1.01	1.298	
9	FRT-PLM	PlumPrune (dried plu	1.31	0.77	12.836	57.883	3.68	4.510	
10	FRT-CHI	Chico fruit	1.75	1.03	6.123	15.144	4.73	2.473	

Copy the result from cell I2 to cells I3:I82. Right-click and format cells with "Number" and three decimal places.

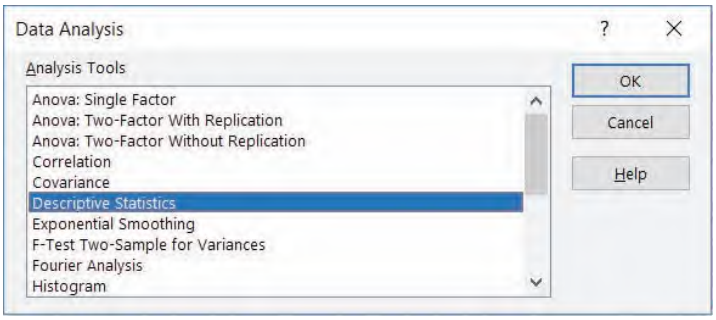
Analysis Task 2: Calculate Descriptive Statistics of Inventory Turnover and Waste.

1. Exotic Fruits wants to hold inventory that is above the median (or mean) inventory turnover ratio, but also that is below the median (or mean) waste as a % of costs. It also wants to know the products with the maximum waste. The greater the inventory turnover, the more efficient are Exotic Fruits' processes. The lower the waste for a particular fruit (as a percent of cost of goods sold), the more profitable that fruit is. To determine the means, medians, and maximums, we will use the Descriptive Statistics option in Excel's Data Analysis ToolPak. (For instructions on how to install this add-in, consult Appendix E at the end of this book.)
2. Choose the Data Analysis Option in **Data> Analysis > Data Analysis**, as shown in Lab Exhibit 4.1.4.

Lab Exhibit 4.1.4

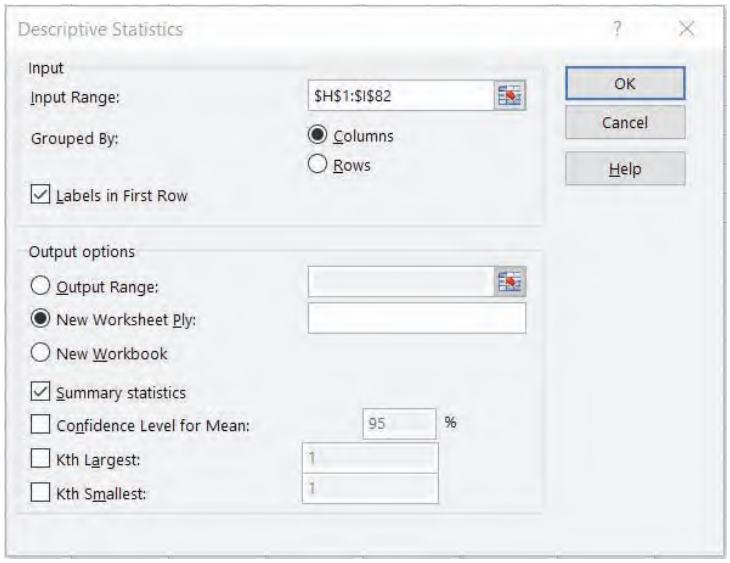


3. Select the Descriptive Statistics Analysis Tool and then click OK, as shown in Lab Exhibit 4.1.5.



Lab Exhibit 4.1.5

4. In the Descriptive Statistics dialog box, insert "$H$1:$I$82". Click Labels in First Row and click Summary Statistics, as shown in Lab Exhibit 4.1.6. Click OK.



Lab Exhibit 4.1.6

5. The resulting descriptive statistics are as shown in Lab Exhibit 4.1.7.

Lab Exhibit 4.1.7

	A	B	C	D
1	*Inventory Turnover*		*Waste as a % of Costs*	
2				
3	Mean	5.661095515	Mean	0.136110836
4	Standard Error	0.45537057	Standard Error	0.012282511
5	Median	4.467292249	Median	0.116705165
6	Mode	#N/A	Mode	#N/A
7	Standard Deviation	4.098335134	Standard Deviation	0.110542603
8	Sample Variance	16.79635087	Sample Variance	0.012219667
9	Kurtosis	3.488759241	Kurtosis	2.926831136
10	Skewness	1.775795275	Skewness	1.511460733
11	Range	20.79597681	Range	0.570726257
12	Minimum	1.280482453	Minimum	0.002081795
13	Maximum	22.07645926	Maximum	0.572808051
14	Sum	458.5487367	Sum	11.02497774
15	Count	81	Count	81

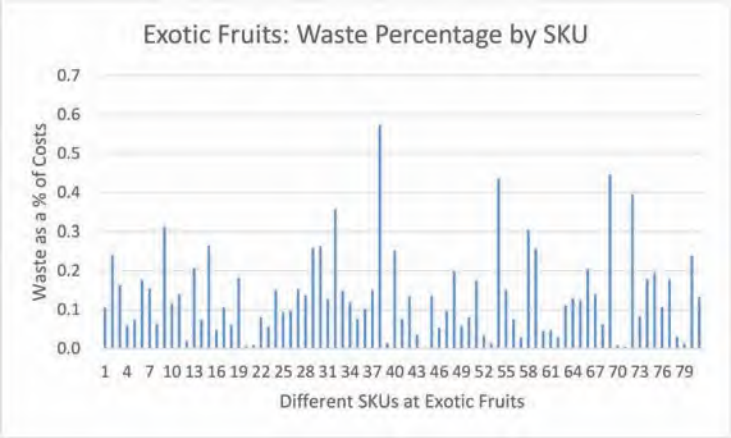


Report the Results

We'll use a visualization to report the results and highlight which types of fruit might require additional investigation. To compare values in the data set, we'll create a column chart (similar to a bar chart). To do so, proceed as follows:

1. Highlight cells I1:I82.
2. In menu, select **Insert > Recommended Charts > All Charts > Column > Clustered Column** and click OK.
3. Add the axis title "Different SKUs at Exotic Fruits" to the horizontal axis and "Waste as a % of Costs" to the vertical axis by clicking the chart, and then click the Chart Design tab. Select **Add Chart Element > Axis Titles**.
4. Add the chart title "Exotic Fruits: Waste Percentage by SKU". To do so, click on the Chart Design tab, then select **Add Chart Element > Chart Title**. The resulting visualization should be as shown in Lab Exhibit 4.1.8.

Lab Exhibit 4.1.8



Exotic Fruits now has the analysis it needs to evaluate its inventory, including means, medians, minimums, and maximums for inventory turnover and waste percentage, along with a visualization of which items create the biggest waste. Exotic Fruits can now make important decisions.

Assessment

1. Take a screenshot of the output for descriptive statistics, paste it into a Word document named "Lab 4.1 Excel Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the visualization, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 4.1 Excel Submission.docx via Connect, if assigned.

Alternate Lab 4.1 Excel

Evaluating Inventory Using Inventory Turnover, Waste, and Profit Margins

Apply the same steps in Lab 4.1 Excel (and continue using Lab 4.1 Data.xlsx), but now compute the profit margin (Sales price per pound minus Cost per pound) and profit margin percentage (margin divided by sales price per pound). Then use Excel's Data Analysis

ToolPak to compute the descriptive statistics (mean, median, maximum, minimum) for the profit margin percentage.

Required

1. Compute the profit margin and **profit margin** percentage ratios.
2. Summarize the profit margin ratios using descriptive statistics.
3. Create a visualization showing the **margin percentage**. Include chart title and axis titles.

Assessment

1. Take a screenshot of the output for descriptive statistics, paste it into a Word document named "Alt Lab 4.1 Excel Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the visualization, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 4.1 Excel Submission.docx via Connect, if assigned.

Profit margin: Total sales less total expenses.

Profit margin percentage: The ratio of a company's profits (total sales less total expenses) to its total sales.



DATA VISUALIZATION

Lab 4.2 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Using Conditional Formatting to Perform a Bank Reconciliation

Keywords

Diagnostic Analytics, General Ledger, Bank Reconciliation, Conditional Formatting

Decision-Making Context

An important internal control for cash is a bank reconciliation, which reconciles the company's cash general ledger account to the company's bank statement. The cash transactions (cash, checks, and other deposits) that are not recorded in both the bank statement and the company's general ledger need to be reconciled.

To perform a reconciliation, the company needs to reconcile the cash balance recorded in its general ledger (GL) with the cash that the bank collected or charged (disbursed) without the company's knowledge. The company also needs to reconcile the cash balance on its bank statement with the transactions recorded in the GL but not known at the bank. Reconciliations help us find recording errors (or possible fraudulent actions) that either the bank or the company made.

conditional formatting: Application of specific formatting to cells (in Excel or Google Sheets) that meet certain criteria.

In this lab, you will use conditional formatting to find items that need to be reconciled. Specifically, you will use conditional formatting to find GL cash transactions that are different from those reported by the bank. In this case, conditional formatting reveals the unique items in the two columns of numbers that will be important to our cash reconciliation.

Required

1. Find the items that need to be reconciled between the company general ledger and the bank statement.

Specify the Question

What differences exist in the cash transactions recorded by the bank and the cash transactions recorded in the company's records?

Obtain the Data

Family First Cabins, a fictitious company, has cash transactions recorded in its general ledger and cash transactions reported on its bank statement. The following key terms will be helpful in understanding a bank reconciliation.

Outstanding checks: Checks written by the company but not yet processed by the bank.
Outstanding deposits: Deposits recorded by the company but not yet processed by the bank.
Non-sufficient funds (NSF) checks: Checks deposited by the company but not paid by the bank because the check writer has insufficient funds.
Electronic funds transfer (EFT, wire transfer): Transfers received by the bank but not yet reflected in the company's records.
Notes and interest: *Notes* are loans made to bank customers and collected by the bank. *Interest* is the cost of borrowing, usually expressed as a percentage of the loan, and payable by the borrower. The company may not be aware of the amounts of these items until it receives the bank statement.

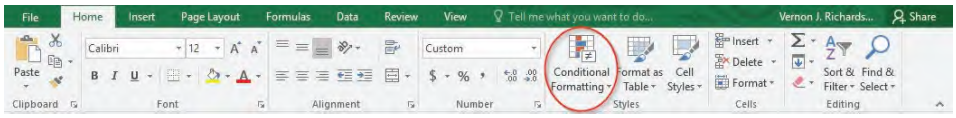
Bank service fees: The bank charges fees for its services—for example, a monthly fee for each checking account. The bank deducts these fees directly from the company's checking account. The company may not account for the fee until it receives the monthly bank statement.
Errors: Sometimes a company or its bank makes a mistake in recording a transaction. Analysts find these errors by comparing the amounts for each transaction on the general ledger and each transaction on the bank statement.

Open Excel file Lab 4.2 Data.xlsx and browse the cash transactions that were recorded in the company's general ledger and the cash transactions that were recorded by the bank.

Analyze the Data

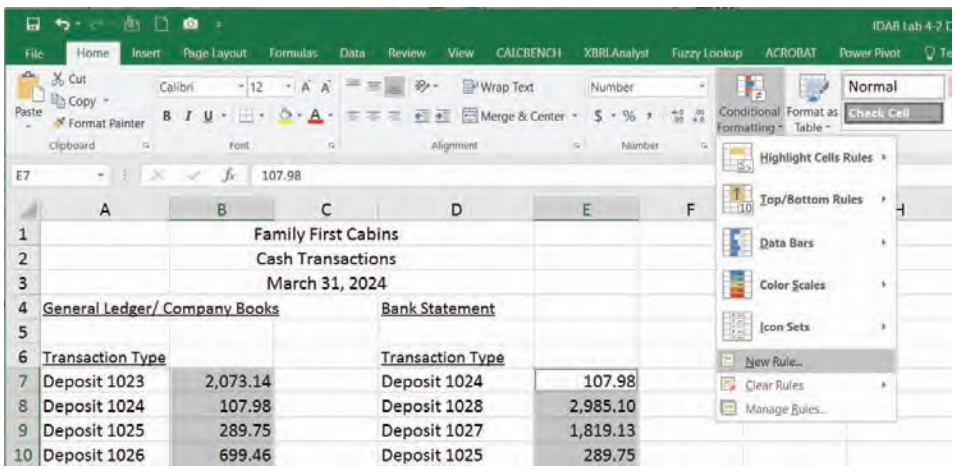
Analysis Task 1: Use Conditional Formatting to Identify Transactions That Do/Do Not Need Reconciliation.

1. Use Excel's conditional formatting function to find the items that need to be reconciled. Conditional formatting applies different formats (or highlights) to cells depending on the content of the individual cells. Use Lab Exhibit 4.2.1 to find the conditional formatting icon.



Lab Exhibit 4.2.1

2. Hold the Control Key and highlight both columns of numbers (in columns B and E). Next, select Conditional Formatting and then click on New Rule (as shown in Lab Exhibit 4.2.2).



Lab Exhibit 4.2.2

3. Select Format only unique or duplicate values and select duplicate under Format all. Click on the Format . . . command to set the color fill you would like for the unique values.

4. Click OK, and the screen will format the unique numbers, highlighting the numbers in each column that need to be reconciled (as shown in Lab Exhibit 4.2.3).

Lab Exhibit 4.2.3

	A	B	C	D	E
1		Family First Cabins			
2		Cash Transactions			
3		March 31, 2024			
4	General Ledger/ Company Books			Bank Statement	
5					
6	Transaction Type			Transaction Type	
7	Deposit 1023	2,073.14		Deposit 1024	107.98
8	Deposit 1024	107.98		Deposit 1028	2,985.10
9	Deposit 1025	289.75		Deposit 1027	1,819.13
10	Deposit 1026	699.46		Deposit 1025	289.75
11	Deposit 1027	1,819.13		Deposit 1026	699.46
12	Deposit 1028	2,985.10		Deposit 1023	2,073.14
13	Deposit 1029	993.09		Deposit 1029	993.09
14	Deposit 1030	3,010.23		Deposit 1030	3,010.23
15	Deposit 1031	1,064.20		Check 11762	1,414.13
16	Deposit 1032	1,065.38		Check 11764	2,041.73
17	Deposit 1033	2,613.29		Check 11765	987.34
18	Check 11762	1,414.13		Check 11766	1,501.19
19	Check 11763	11,783.98		Check 11767	2,278.18
20	Check 11764	2,041.73		Check 11768	4,398.01
21	Check 11765	987.34		Check 11769	1,349.56
22	Check 11766	1,521.19		Check 11770	2,430.61
23	Check 11767	2,278.18		Note Collect 1	1,500.00
24	Check 11768	4,398.01		Interest Earned	11.05
25	Check 11769	1,349.56		NSF Check 1013	2,645.05
26	Check 11770	2,430.61		NSF Check 1017	924.12
27	Check 11771	1,513.19		NSF Fees 1	25.00
28	Check 11772	2,985.09		NSF Fees 2	65.00
29	Check 11773	865.23			
30	Check 11774	152.90			

Analysis Task 2: Analyze the Different Items Needing Reconciliation.

1. The cells highlighted in blue represent transactions that have been recorded both in the general ledger and by the bank (example: Deposit 1029).
2. In contrast, the cells not highlighted in blue are the items that need to be reconciled (example: Check 11763).
3. The cells in the general ledger column not highlighted in blue are items recorded by the company but not by the bank. The checks not highlighted in blue are considered outstanding checks (example: Check 11774).
4. The deposits not highlighted in blue in the general ledger column are considered outstanding deposits (example: Deposit 1031).
5. The cells in the bank statement column not highlighted in blue are items recorded by the bank but not by the company (example: NSF Fees 1).
6. The cells pertaining to the same check or deposit number not highlighted in both the general ledger and the bank statement column are considered to be errors. For example, see Check 11766, where the company (on its general ledger) and the bank recorded different amounts.

Report the Results



Conditional formatting is a natural way to visualize the items that need and do not need reconciling.

Bank reconciliations, which identify anomalies and outliers, are a good example of diagnostic analytics. Bank reconciliations are a basic internal control that companies use to ensure their cash transactions are properly recorded. By comparing the cash transactions recorded in the general ledger to the independent records of a bank, the company and its auditors can be more certain that its cash transactions are properly recorded. If transactions were not properly recorded, the company and its auditors can identify and correct those errors. The company may also use the bank reconciliation to identify bank errors.

Assessment

1. Take a screenshot of your final bank reconciliation, paste it into a Word document named "Lab 4.2 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Lab 4.2 Excel Submission.docx via Connect, if assigned.

Alternate Lab 4.2 Excel

Using Conditional Formatting to Perform a Bank Reconciliation

Apply the same steps in Lab 4.2 Excel to the Alt Lab 4.2 Data.xlsx data set, with a completed reconciliation. Open Excel file Alt Lab 4.2 Data.xlsx.

Assessment

1. Take a screenshot of your final bank reconciliation, paste it into a Word document named "Alt Lab 4.2 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 4.2 Excel Submission.docx via Connect, if assigned.

Lab 4.3 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Applying Benford's Law

Keywords

Financial Accounting, Benford's Law, Abnormal Transactions

Decision-Making Context

An important role of a company's internal auditor or fraud examiners is to detect fraudulent and suspicious transactions. While traditional audit approaches randomly select samples from daily transactions, diagnostic data analytics works to find anomalies to help auditors focus their efforts on the more critical, possibly fraudulent accounts. To show the breadth of naturally occurring populations, we demonstrate the use of Benford's Law using the population of countries of the world.

Required

Test the population of countries of the world for conformity with Benford's Law.

Specify the Question

Is the frequency of first digits in the set of numbers representing country populations different from expectations set by Benford's Law?

Obtain the Data

Open Lab 4.3 Data.xlsx. The data come from a Worldometer page titled "Countries in the World by Population (2020)," [5] which contains population data for over 200 countries. The data dictionary is below.

Data Dictionary

#:	Ranking by country population
Country (or Dependency):	Country name
Population-2020:	Population of country in 2020 (estimated)
Yearly Change:	Percentage change in annual population
Net Change:	Change in annual population

Analyze the Data

Analysis Task 1: Identify the First Digit for the Population of Each Country.

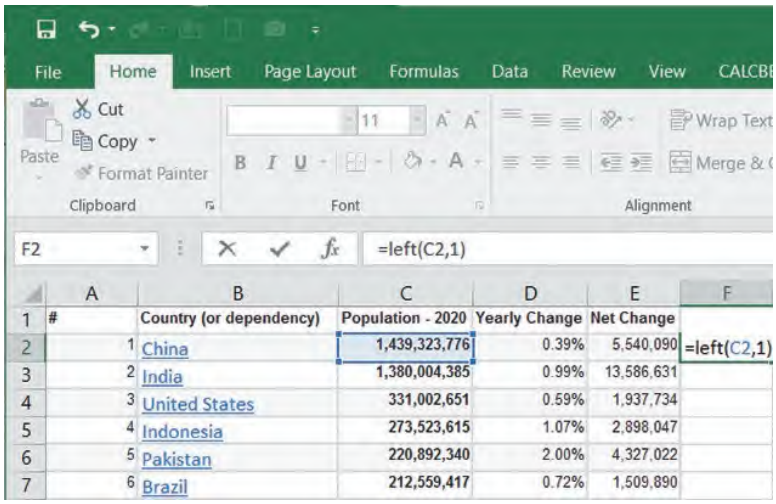
1. Browse the spreadsheet very quickly to make sure there aren't any obvious errors from opening the Excel file.
2. Double-click on the Sheet Tab and rename the new worksheet tab "Data" to replace "Sheet1" (as shown in Lab Exhibit 4.3.1).

Lab Exhibit 4.3.1



[5] https://www.worldometers.info/world-population/population-by-country/. Accessed April 6, 2020.

3. In cell F1, type the words “First Digit”. In cell F2, insert the formula “=left(C2,1)”, as shown in Lab Exhibit 4.3.2, and hit Enter. Because Benford’s Law is a law regarding the distribution of the first digits, the =left() formula in Excel will tell us the first digit. The “1” entered in the formula suggests we want only one digit: the first digit.



	A	B	C	D	E	F
1	#	Country (or dependency)	Population - 2020	Yearly Change	Net Change	
2	1	China	1,439,323,776	0.39%	5,540,090	=left(C2,1)
3	2	India	1,380,004,385	0.99%	13,586,631	
4	3	United States	331,002,651	0.59%	1,937,734	
5	4	Indonesia	273,523,615	1.07%	2,898,047	
6	5	Pakistan	220,892,340	2.00%	4,327,022	
7	6	Brazil	212,559,417	0.72%	1,509,890	

Lab Exhibit 4.3.2

4. Copy the result down from cell C2 to cells C3:C236. The result will look as shown in Lab Exhibit 4.3.3.

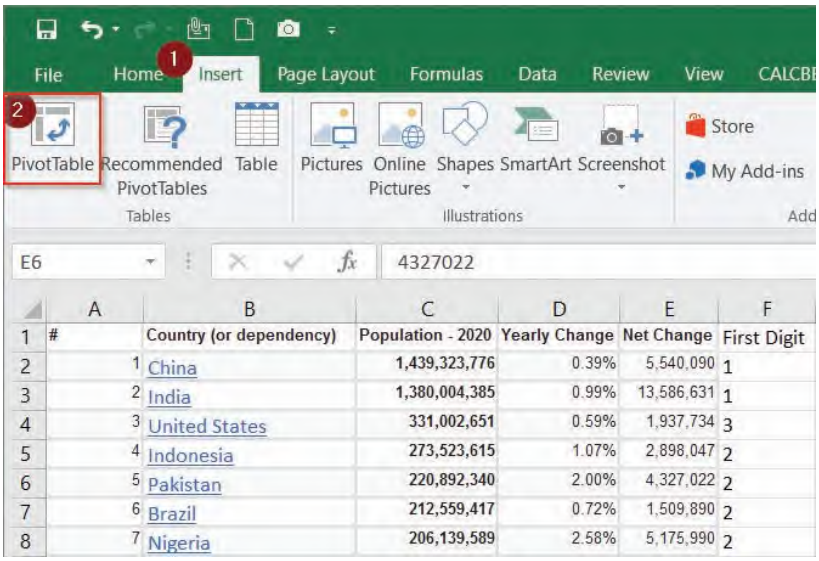
	A	B	C	D	E	F
1	#	Country (or dependency)	Population - 2020	Yearly Change	Net Change	First Digit
2	1	China	1,439,323,776	0.39%	5,540,090	1
3	2	India	1,380,004,385	0.99%	13,586,631	1
4	3	United States	331,002,651	0.59%	1,937,734	3
5	4	Indonesia	273,523,615	1.07%	2,898,047	2
6	5	Pakistan	220,892,340	2.00%	4,327,022	2
7	6	Brazil	212,559,417	0.72%	1,509,890	2
8	7	Nigeria	206,139,589	2.58%	5,175,990	2
9	8	Bangladesh	164,689,383	1.01%	1,643,222	1
10	9	Russia	145,934,462	0.04%	62,206	1
11	10	Mexico	128,932,753	1.06%	1,357,224	1
12	11	Japan	126,476,461	-0.30%	-383,840	1
13	12	Ethiopia	114,963,588	2.57%	2,884,858	1
14	13	Philippines	109,581,078	1.35%	1,464,463	1
15	14	Egypt	102,334,404	1.94%	1,946,331	1
16	15	Vietnam	97,338,579	0.91%	876,473	9

Lab Exhibit 4.3.3

Analysis Task 2: Create a Count of the First Digit for the Data Set.

1. The next step is to get a count, or frequency, for each digit. To do so, we use an Excel pivot table. Select **Insert > PivotTable**, as shown in Lab Exhibit 4.3.4.

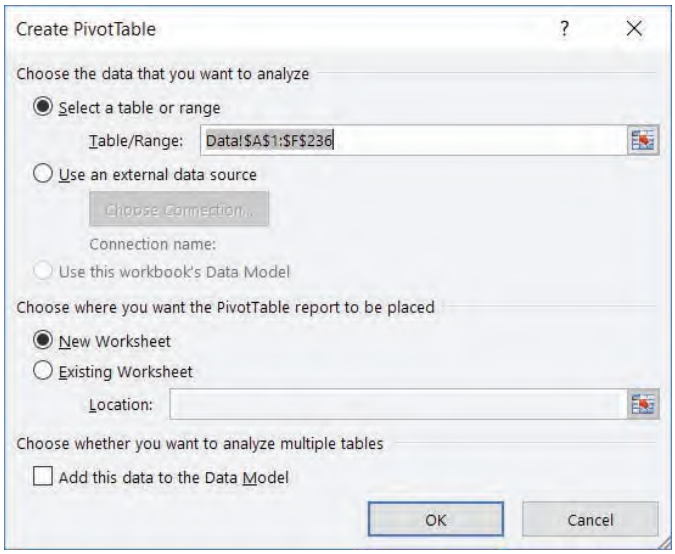
Lab Exhibit 4.3.4



	A	B	C	D	E	F
1	#	Country (or dependency)	Population - 2020	Yearly Change	Net Change	First Digit
2	1	China	1,439,323,776	0.39%	5,540,090	1
3	2	India	1,380,004,385	0.99%	13,586,631	1
4	3	United States	331,002,651	0.59%	1,937,734	3
5	4	Indonesia	273,523,615	1.07%	2,898,047	2
6	5	Pakistan	220,892,340	2.00%	4,327,022	2
7	6	Brazil	212,559,417	0.72%	1,509,890	2
8	7	Nigeria	206,139,589	2.58%	5,175,990	2

2. In creating the pivot table, select the Table/Range from $A$1:$F$236 (or more specifically Data!$A$1:$F$236, as shown in Lab Exhibit 4.3.5).

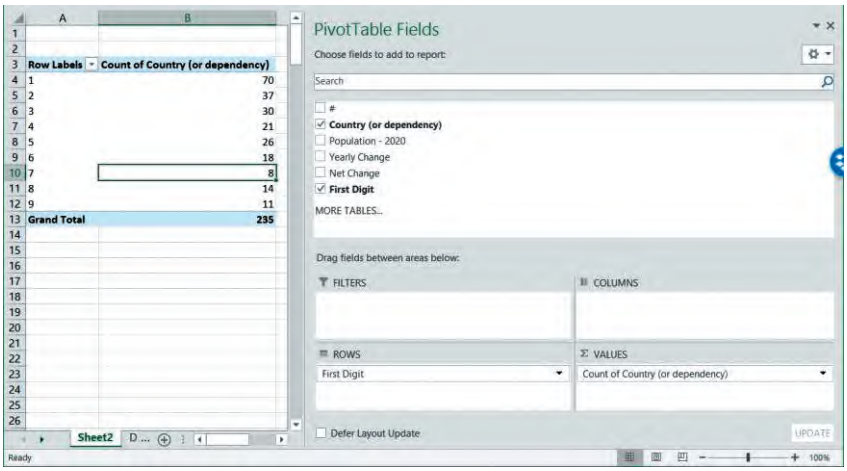
Lab Exhibit 4.3.5



Inputs into Pivot Table
Columns:
Rows: [First Digit]
ΣValues: [Country (or dependency]

3. Drag [First Digit] from **FIELD NAME** into the **ROWS** and [Country (or dependency)] from **FIELD NAME** into **ΣVALUES** field in the pivot table. The ΣVALUES will default to "Count of Country (or dependency)". The resulting pivot table will look as shown in Lab Exhibit 4.3.6.

Lab Exhibit 4.3.6



Analysis Task 3: Compare the Distribution of First Digits to the Expected Distribution of First Digits Using Benford's Law.

1. Copy the results of the pivot table (cells A4:B13) to cells A2:B11, using Paste Special → Values, in another sheet. Title the sheet "Benford's Law" (as shown in Lab Exhibit 4.3.7).

Lab Exhibit 4.3.7

	A	B
1	First_Digit	Count
2	1	70
3	2	37
4	3	30
5	4	21
6	5	26
7	6	18
8	7	8
9	8	14
10	9	11
11	Grand Total	235
12		

Benford's Law

2. The next step is to calculate the percentage of each digit and compare it with expected percentage of Benford's Law. In cell A1, type "First Digit"; in cell B1, type "Count"; in cell C1, type "Percentage". In cell C2, divide the count of first-digit numbers by the total

number of transactions in B11. Note that we use an absolute reference for $B$11 so that we can copy the formula down. Use the formula to calculate the percentage of first-digit 1, as shown in Lab Exhibit 4.3.8. Then copy this formula all the way down to cell C10.

Lab Exhibit 4.3.8

	A	B	C
1	First_Digit	Count	Percentage
2	1	70	=B2/$B$11
3	2	37	
4	3	30	
5	4	21	
6	5	26	
7	6	18	
8	7	8	
9	8	14	
10	9	11	
11	**Grand Total**	235	

3. In cell D1, type "Benford's Law". In cell D2, calculate the expected frequency based on Benford's Law using the formula "=LOG10(1/A2+1)", as shown in Lab Exhibit 4.3.9.

Lab Exhibit 4.3.9

	A	B	C	D
1	First_Digit	Count	Percentage	Benford's Law
2	1	70	0.29787234	=LOG10(1/A2+1)
3	2	37	0.157446809	
4	3	30	0.127659574	
5	4	21	0.089361702	
6	5	26	0.110638298	
7	6	18	0.076595745	
8	7	8	0.034042553	
9	8	14	0.059574468	
10	9	11	0.046808511	
11	**Grand Total**	235		
12				
13				

4. Copy the formula to the last row of First_Digit, as shown in Lab Exhibit 4.3.10.

Lab Exhibit 4.3.10

D2 =LOG10(1/A2+1)

	A	B	C	D	E
1	First_Digit	Count	Percentage	Benford's Law	
2	1	70	0.29787234	0.301029996	
3	2	37	0.157446809	0.176091259	
4	3	30	0.127659574	0.124938737	
5	4	21	0.089361702	0.096910013	
6	5	26	0.110638298	0.079181246	
7	6	18	0.076595745	0.06694679	
8	7	8	0.034042553	0.057991947	
9	8	14	0.059574468	0.051152522	
10	9	11	0.046808511	0.045757491	
11	**Grand Total**	235			
12					

Sheet2 | Benford's Law | Data

Ready

5. Format the Percentage and Benford's Law columns as percentages. (*Hint:* Select columns of Percentage and Benford's Law by selecting **Home > Number > Percentage**.) The resulting spreadsheet should appear as shown in Lab Exhibit 4.3.11.

Lab Exhibit 4.3.11

	A	B	C	D	E
1	First_Digit	Count	Percentage	Benford's Law	
2	1	70	30%	30%	
3	2	37	16%	18%	
4	3	30	13%	12%	
5	4	21	9%	10%	
6	5	26	11%	8%	
7	6	18	8%	7%	
8	7	8	3%	6%	
9	8	14	6%	5%	
10	9	11	5%	5%	
11	**Grand Total**	**235**			

6. Calculate the difference between the expected percentage of first digits based on Benford's Law and the percentage of our data set (after copying down the column for all of the First_Digit rows), as shown in Lab Exhibit 4.3.12.

Lab Exhibit 4.3.12

MATCH ⋮ ✕ ✓ f_x =C2-D2

	A	B	C	D	E
1	First_Digit	Count	Percentage	Benford's Law	Difference
2	1	70	30%	30%	=C2-D2
3	2	37	16%	18%	
4	3	30	13%	12%	
5	4	21	9%	10%	
6	5	26	11%	8%	
7	6	18	8%	7%	
8	7	8	3%	6%	
9	8	14	6%	5%	
10	9	11	5%	5%	
11	**Grand Total**	**235**			

Analysis Task 4: Compare the Actual Percentage to the Expected Percentage of First Digits Using Statistics.

1. Next, we will compare the distribution of the first digits in our data set with the expected values based on Benford's Law.
2. We will also use a statistical test, the chi-square, to test if the distribution of first digits of our data set is significantly different from that expected by Benford's Law.

We have two hypotheses (the null and the alternate):
H_0: Our data set's first digits conform to Benford's Law.
H_1: Our data set's first digits do not conform to Benford's Law.

When the null hypothesis is rejected (p-value <=0.05, or alpha), additional investigations are needed to test for fraud or irregularities.

3. Add an additional column labeled "Expected Count". Expected Count equals D2* $B$11 or the Benford's percentage multiplied by the total amount of transactions. Copy down the column. (Note: Round to the nearest whole number.) See Lab Exhibit 4.3.13.

Lab Exhibit 4.3.13

	A	B	C	D	E	F
1	First_Digit	Count	Percentage	Benford's Law	Difference	Expected Count
2	1	70	30%	30%	0%	=D2*$B$11
3	2	37	16%	18%	-2%	41
4	3	30	13%	12%	0%	29
5	4	21	9%	10%	-1%	23
6	5	26	11%	8%	3%	19
7	6	18	8%	7%	1%	16
8	7	8	3%	6%	-2%	14
9	8	14	6%	5%	1%	12
10	9	11	5%	5%	0%	11
11	**Grand Total**	**235**				

4. Use the chi-square test function to determine the *p*-value of the statistical test to determine if the actual distribution of first digits is statistically different than that expected by Benford's Law. In cell B13, type "chi-square test". In cell C14, insert the equation "=CHISQ.TEST(B2:B10,F2:F10)" and hit enter.
   Note the screenshots in Lab Exhibits 4.3.14 and 4.3.15. If the resulting *p*-value of the chi-square test is less than or equal to 0.05, then we will reject the null hypothesis that the distributions are not different from each other and investigate the reasons the distributions are different from each other.

Lab Exhibit 4.3.14

C13 =CHISQ.TEST(B2:B10,F2:F10)

	A	B	C	D	E	F	G
1	First_Digit	Count	Percentage	Benford's Law	Difference	Expected Count	
2	1	70	30%	30%	0%	71	
3	2	37	16%	18%	-2%	41	
4	3	30	13%	12%	0%	29	
5	4	21	9%	10%	-1%	23	
6	5	26	11%	8%	3%	19	
7	6	18	8%	7%	1%	16	
8	7	8	3%	6%	-2%	14	
9	8	14	6%	5%	1%	12	
10	9	11	5%	5%	0%	11	
11	**Grand Total**	**235**					
12							
13		chi-square test	=CHISQ.TEST(B2:B10,F2:F10)				
14							

Lab Exhibit 4.3.15

	A	B	C	D	E	F
1	First_Digit	Count	Percentage	Benford's Law	Difference	Expected Count
2	1	70	30%	30%	0%	71
3	2	37	16%	18%	-2%	41
4	3	30	13%	12%	0%	29
5	4	21	9%	10%	-1%	23
6	5	26	11%	8%	3%	19
7	6	18	8%	7%	1%	16
8	7	8	3%	6%	-2%	14
9	8	14	6%	5%	1%	12
10	9	11	5%	5%	0%	11
11	**Grand Total**	**235**				
12						
13		chi-square test	0.59			

5. The *p*-value for the chi-square statistic is 0.59 (rounded). Thus, we do not reject H_0, because the *p*-value is greater than our alpha of 0.05. Our results indicate that, while the actual first-digit distribution is not exactly the same as the expected distribution of Benford's Law, the differences are not statistically different from each other.

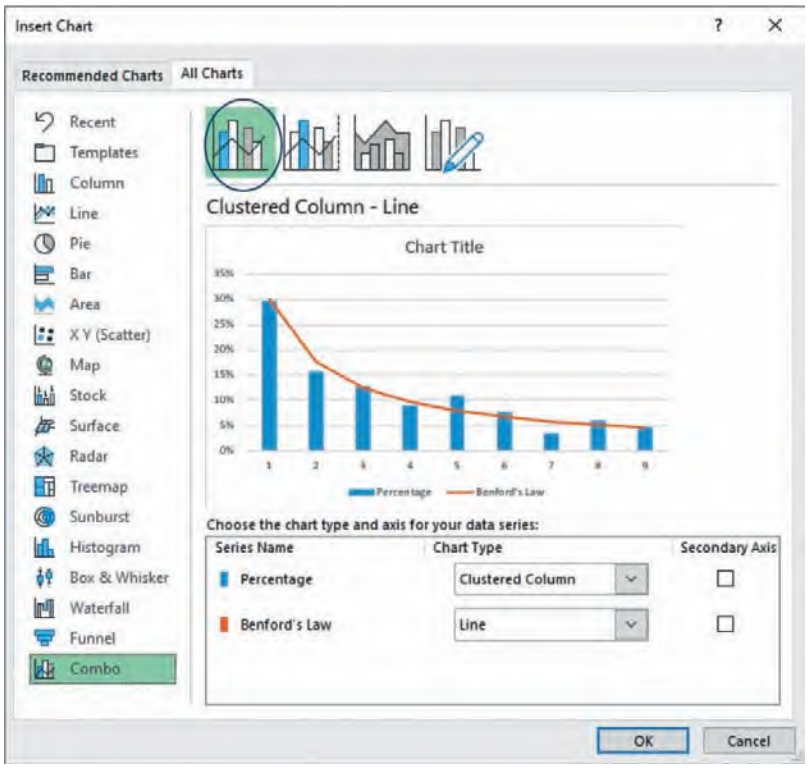
Report the Results


DATA VISUALIZATION

Often the best way to report the results is to provide a visualization.

1. To view the differences between the actual percentage (from column C, Percentage) and the expected percentage (from column D, Benford's Law), insert a combo chart.
   a. Highlight the Percentage and Benford's Law columns, including the headers.
   b. Select **Insert > Charts > Insert Combo Chart**, as shown in Lab Exhibit 4.3.16.

Lab Exhibit 4.3.16

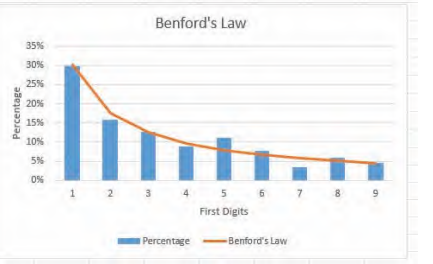


   c. Change the name of the chart to Benford's Law.
   d. Add the axis title "First Digits" to the horizontal axis and "Percentage" to the vertical axis by clicking the chart, and then clicking the Chart Design tab. Select **Add Chart Element > Axis Titles.**

The graph in Lab Exhibit 4.3.17 suggests that fewer-than-expected populations started with the number 2 or 7, but populations starting with the number 5 occurred more frequently than predicted by Benford's Law. If this had been a distribution of sales or payment transactions (rather country populations), internal auditors or fraud examiners might focus their attention on transactions beginning with numbers 2, 5, and 7 to see why they might be different than expected.

Lab Exhibit 4.3.17

	A	B	C	D	E	F
1	First_Digit	Count	Percentage	Benford's Law	Difference	Expected Count
2	1	70	30%	30%	0%	71
3	2	37	16%	18%	-2%	41
4	3	30	13%	12%	0%	29
5	4	21	9%	10%	-1%	23
6	5	26	11%	8%	3%	19
7	6	18	8%	7%	1%	16
8	7	8	3%	6%	-2%	14
9	8	14	6%	5%	1%	12
10	9	11	5%	5%	0%	11
11	Grand Total	235				
12						
13		chi-square test	0.59			



In sum, Benford's Law can help internal auditors or fraud examiners identify fraud and/or errors by finding first digits of numbers that are different from expectations. Benford's Law helps identify specific samples, accounts, or transactions that require additional verification, increasing work efficiency and effectiveness.

Assessment

1. Take a screenshot of your worksheet titled "Benford's Law", paste it into a Word document named "Lab 4.3 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Lab 4.3 Excel Submission.docx via Connect, if assigned.

Alternate Lab 4.3 Excel

Applying Benford's Law

Apply the same steps in Lab 4.3 Excel to the Alt Lab 4.3 Data.xlsx data set, with Benford's Law analysis.

Specify the Question

Is the frequency of first digits in the set of numbers representing country populations different from expectations set by Benford's Law?

This data set represents the population of the countries of the world in 1950.[6]

Open Excel file Alt Lab 4.3 Data.xlsx.

Assessment

1. Take a screenshot of your worksheet titled "Benford's Law", paste it into a Word document named "Alt Lab 4.3 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 4.3 Excel Submission.docx via Connect, if assigned.

[6]UN Populations Division, https://photius.com/rankings/world2050_rank.html. Accessed April 2021.

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Chapter 5

Analyze the Data:

Confirmatory Business Analytics (Predictive Analytics and Prescriptive Analytics)

A Look Back

- The third step of the SOAR model is ANALYZE THE DATA.
- Chapter 4 introduced five types of business analytics, emphasizing descriptive and diagnostic analytics (also called exploratory business analytics).
- Descriptive and diagnostic analytics use specific analysis techniques.

A Look at This Chapter

- This chapter emphasizes predictive and prescriptive analytics (also called confirmatory business analytics).
- Predictive analytics techniques help a company forecast certain events and outcomes. This chapter emphasizes classification, regression, and time series analysis.
- The prescriptive analytics techniques explained in this chapter help a company respond to what it believes will happen. These techniques include what-if analysis, cash flow analysis, marginal analysis, goal-seek analysis, and what-if scenarios.

A Look Ahead

- The fourth and final step of the SOAR model is REPORT THE RESULTS.
- Chapter 6 emphasizes matching the report type (such as written reports and data visualizations) to the information needed to answer decision-makers' questions.
- Different types of data visualizations are used to accomplish different purposes.



Travel mania/Shutterstock

COVID-19 and Panic Buying

Thanks to modern technologies, the supply chain has become extremely efficient. Using IoT (Internet of Things) and RFID (Radio Frequency ID Tags), retailers and distributors know where all their products are at any given time. Using predictive analytics, they can predict what will sell quickly and what will not. Forecasting and delivery models allow retailers to keep their shelves stocked until the next shipment arrives.

Given the tight integration of the supply chain from manufacturing to retailer to the end consumer, any unforeseen blip, however small, causes a significant ripple throughout the supply chain. COVID-19 was not a small blip in the system; it was a major blip in the system. People bought hundreds of rolls of toilet paper instead of the much more common 4-pack, creating major inventory shortages. The cycle became self-perpetuating as more people began to worry that important "stuff" would soon be unavailable, causing major shortages of tissues, paper towels, bleach, and other disinfectants.

We use predictive analytics to predict the impact of small and large blips on operations and the supply chain, allowing us to prevent many (but not all) problems before they arise.

Source: Rob Armstrong, "Don't Let Panic Worsen the COVID-19 Crisis: Let Data Run the Supply Chain," by Rob Teradata, https://www.teradata.com/Blogs/Don%E2%80%99t-let-panic-worsen-the-COVID-19-crisis-Let-data-run-the-supply-chain. Accessed September 2021.

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

- **LO 5.1** Define *confirmatory analytics*, including predictive and prescriptive analytics.
- **LO 5.2** Describe predictive analytics and techniques, including classification, regression, and time series analysis.
- **LO 5.3** Define *base rates* and *base rate fallacy*.
- **LO 5.4** Analyze data to determine the best course of action using prescriptive analytics and its techniques.
- **LO 5.5** Summarize the different types of data analytics and the business questions they attempt to answer.

LABS FOR CHAPTER 5

- **LAB 5.1** **Excel:** Evaluating the Relationship Between Sales and R&D Expenditures
- **LAB 5.2** **Excel:** Evaluating the Relationship Between Sales and R&D Expenditures: Testing for a Nonlinear Relationship
- **LAB 5.3** **Excel:** Forecasting Product Demand Using Time Series Analysis
- **LAB 5.4** **Tableau:** Forecasting Product Demand Using Time Series Analysis
- **LAB 5.5** **Power BI:** Forecasting Product Demand Using Time Series Analysis
- **LAB 5.6** **Excel:** Forecasting Product Demand Using Regression
- **LAB 5.7** **Excel:** Assessing the Returns to Education: To MBA or Not to MBA?
- **LAB 5.8** **Excel:** Applying Scenario Analysis: Possible Trade War

LO 5.1

Define *confirmatory analytics*, including predictive and prescriptive analytics.

5.1 CONFIRMATORY DATA ANALYTICS

Exploratory data analytics are the initial descriptive and diagnostic data analytics that help us summarize and explain performance. They explore the data and allow us to generate questions and hypotheses that need further explanation. To address those questions and hypotheses, we use **confirmatory data analytics**, which are predictive and prescriptive analytics that use various techniques and statistics to evaluate the likelihood (probability) of a future event or outcome occurring. While it is not necessary to perform exploratory data analytics (descriptive and diagnostic analytics) to understand the past before we conduct confirmatory data analytics, analytics are usually performed in this order.

Exploratory data analytics (covered in Chapter 4) generates *deterministic* output, which is an analysis where the same data input always creates the same data output. Deterministic data are known with certainty. In contrast, confirmatory data analytics provides less deterministic output and more *probabilistic* models, judging the likelihood of a future event or outcome occurring. Probabilistic data are not known beforehand and are subject to chance. With probabilistic models, there is always an element of uncertainty regarding whether or not the prediction will actually happen.

PROGRESS CHECK

1. Is the amount of money that a company currently has in its bank account deterministic or probabilistic? Is the chance that a company will go bankrupt deterministic or probabilistic? Explain.
2. Is predicting whether it will rain tomorrow or next week probabilistic or deterministic? Explain.

LO 5.2

Describe predictive analytics and techniques, including classification, regression, and time series analysis.

5.2 PREDICTIVE ANALYTICS

Predictive analysis addresses three main questions: (1) "Will it happen in the future?" (2) "What is the probability something will happen?" and (3) "Can we forecast what will happen?" **Predictive analytics** provides foresight by using statistical techniques to identify patterns in historical data and assess likelihood or probability. For example, a company may use predictive analytics to predict sales with enough accuracy to produce just enough product to meet customer demand without overproducing.

Three common predictive analytics techniques used to answer the question "Will it happen in the future?" are:

1. Classification
2. Regression
3. Forecasting with time series analysis

Exhibit 5.1 summarizes these predictive analytics techniques, along with example questions and specific techniques employed. We describe each technique in more detail in the following sections.

Classification

Classification is a predictive analytics technique to separate or classify a sample (or population) into two or more groups or classes. Classification might be used to identify customer segments, predict whether a company is expected to go bankrupt or not, or determine whether a bank (such as **JP Morgan Chase**) will extend or not extend a loan to a potential borrower.

Predictive Analytics Technique	Example Business Question	Specific Technique Employed
Classification: Separates or classifies a sample (or population) into two or more groups or classes	How do we identify the various customer segments at **Best Buy**?	Run a classification analysis to predict the most appropriate way to segment customers.
Regression: Predicts a specific outcome value based on independent variable inputs	Can we predict the demand for **Samsung** phones based on specific variables that might explain it?	Run a regression analysis to assess the relationship between phone demand and various factors/independent variables (see Lab 5.6 Excel for an example).
Forecasting with Time Series Analysis Predicts future values based on past values of the same variable	Can we predict the demand for **Google** Pixel phones using past demand for the product?	Use time series analysis incorporating past values of phone demand to predict next month's product demand (see Lab 5.3 Excel, Lab 5.4 Tableau, and Lab 5.5 Power BI for examples).

Exhibit 5.1
Predictive Analytics Techniques, Example Business Questions, and Specific Techniques Employed

Note that a classification includes just a few classes, such as customer type 1/customer type 2/customer type 3; go bankrupt/do not go bankrupt; or loan approved/not approved.

Let's now look more closely at a few examples of classification.

Classification: Identifying and Marketing to Customer Segments

Rather than selling to each of its customers individually, the marketing department often uses classification to cluster its customers into similar customer segments (groups or classes), which can allow for more effective targeting of customer segments in marketing and advertising campaigns. Classifying (or clustering) customers into segments is often based on (1) customer touchpoint engagement or (2) customer purchase patterns.

As data are increasingly used to classify customers into different groups, the demand for the ethical collection and use of customer data will continue to intensify. To win and maintain customers' trust, companies must continue to emphasize customer data ethics and demonstrate their commitment to go beyond just legal compliance.[1]



Customer Segmentation Based on Touchpoints

Customer touchpoints refer to the ways in which a company engages with its customers. Four common touchpoints are:

1. Communication touchpoints (websites, email, snail mail, SMS messaging/texting)
2. Human touchpoints (interaction with a salesperson or customer service representative)
3. Physical touchpoints (physical store location, product and packaging)
4. Sensory touchpoints (fragrances, store lighting, store design)

Using predictive analytics, we can classify customers into different segments based on any (or several) of these touchpoints. For example, we can classify customers based on whether they usually evaluate products online instead of at a physical store, and we can then market differently to customers based on this criterion.

[1]Dylan Bushell-Embling, "Data Ethics to Shake Up Digital Marketing," https://www.technologydecisions.com.au/content/it-management/news/data-ethics-to-shake-up-digital-marketing-1015966134. Accessed March 2022.

Customer Segmentation Based on Purchase Patterns

Firms also classify their customers according to what they have bought, or **customer purchase patterns**. For example, a company may segment its customers in terms of their level of sales (customers who bought the most product) or their types of purchases (customers who consistently bought products with the biggest profit margins).

This type of segmentation, along with even more finely classified micro-segmentation, is a rising trend in marketing. Pass Micro-segmentation classifies people into more precise categories, often according to their behavioral intentions, allowing a company to market to the preferences of each customer group, even the smallest of those groups.[2]

MINI CASE: Customer Segmentation at Best Buy

In 2017, **Best Buy** used data to determine that its target customers were fans of technology who like to touch products before buying them ("high-touch" customers). However, as sales moved more online, where in-store high-touch opportunities aren't as possible, Best Buy had to change how it defined its customers. Using its algorithms to segment customers, Best Buy began to search for its target customers across multiple sales channels (online, in store, kiosks, and so forth). Once it identified these target customers, Best Buy tried to offer them high-touch opportunities, including in-store, in-home, and virtual consultations.[3]

Decision-Making with Data

How would Best Buy know if its strategy is working? What data will it use to test its success? How will Best Buy analyze those data?

Classification: Deciding to Extend Loans to Potential Borrowers

Banks evaluate the creditworthiness of companies and individuals to decide whether they will extend a loan to them. Banks use various borrower characteristics to determine whether the customer is a good credit risk—that is, to predict whether the customer is likely to pay back the loan and the interest due. To make this decision, banks use not only the customer's **credit score** (a statistic that evaluates a customer's creditworthiness and ability to repay a loan) but also the customer's work history, level of existing debt, debt-to-income ratio, and home ownership status. They then classify their customers into two groups: those who are eligible for a loan and those who are not.

Classification: Identifying Potentially Fraudulent Financial Statements

Fraud is deception intended to result in financial gain. The types of fraud are virtually limitless; some common methods of committing fraud include reporting fictitious sales and underreporting expenses to make a company appear more profitable than it actually is. We can use classification to classify firms into two groups: (1) those where we expect to

[2]https://www.kdnuggets.com/2019/11/top-8-data-science-use-cases-marketing.html.
[3]D. Berthiaume, "Best Buy Makes It Personal with Google Cloud," 2020, https://chainstoreage.com/best-buy-makes-it-personal-google-cloud; S. Lauchlan, "No One's Amazon-Proof, But Best Buy Is Doing Its Best with a Smart Home Focus," 2017, https://diginomica.com/no-ones-amazon-proof-best-buy-best-smart-home-focus. Accessed September 2021.

find fraud in the financial statements and (2) those where we expect to find no fraud in the financial statements. As an example, consider a company where managers receive a bonus equal to 100 percent of their annual salary if they achieve a certain earnings target. Such a compensation plan raises the likelihood of fraudulent financial statements.

Auditors and investors want to know if there is a high likelihood that a company's reported earnings are fraudulent. Savvy auditors know that there is a greater likelihood of fraud if earnings grew unexpectedly over the past two years, if the board of directors is not truly independent of management, and/or if management has unusually strong incentives to hit certain earnings targets. A classification activity might help them predict which financial statements (or which firms) are likely engaging in fraud and which ones are not.

MINI CASE: Using Technology to Detect Customer Fraud

The Chinese e-commerce company **Alibaba** is perhaps the biggest online commerce company in the world. Using its three main websites, Taobao.com, Tmall.com, and Alibaba.com, it hosts millions of businesses and hundreds of millions of users. It racked up $717 billion in sales in 2021—more than **eBay** and **Amazon** combined. With so many transactions and so many users, Alibaba has worked to capture fraud signals directly from its extensive database of user behaviors and its network. It analyzes those signals in real time and uses machine learning to sort the bad users from the good ones. Specifically, Alibaba has developed five stages of fraud detection for each user: (1) Account Check, (2) Device Check, (3) Activity Check, (4) Risk Strategy, and (5) Manual Review. All of these stages combine to develop a risk score for each user. These fraud risk prevention scores are so valuable that Alibaba shares them with and sells them to other businesses.

Decision-Making with Data

How will Alibaba know if its fraud identification is successful or whether it removes honest individuals from its customer lists?

Sources: Alibaba Financial Statements, 2021; J. Chen, Y. Tao, H. Wang, and T. Chen, "Big Data Based Fraud Risk Management at Alibaba," *Journal of Finance and Data Science, 1*(1), 1-10; Pal, K., "How to Combat Financial Fraud by Using Big Data," 2016, http://www.kdnuggets.com/2016/03/combat-financial-fraud-using-big-data.html.

Regression

Regression analysis is a method of estimating an outcome based on variable inputs. Unlike classification, which groups a variable into just a few categories (for example, fraud or no fraud; extend loan or do not extend loan), regressions typically involve many possible outcomes.

Regression analysis uses mathematics and statistics to determine which variables are associated with an outcome and the extent to which those variables affect the outcome. Regression analysis helps us answer the questions "Which variable inputs matter most in predicting the outcome?" and "Which factors can we ignore?"

Unlike classification, which predicts which class or group an item belongs in, a regression analysis results in numerical output with many possible outcomes, such as estimated future sales dollars.

Let's now examine a few examples to show how regression works.

> **LAB CONNECTION**
>
> **LAB 5.1 EXCEL** and **LAB 5.2 EXCEL** provide a step-by-step discussion of and practice with predicting sales given the amount spent on research and development (typically called R&D or R&D expenditures). Research and development can help to identify new products that customers want to buy or to make the organization more efficient.

Regression: Predicting Product Demand

Determining product demand is important for operations (in particular, for the supply chain) and for budgeting. In an operations setting, product demand is the dependent variable or outcome, and predicting it may require considering various inputs, such as weather conditions, economic conditions, holidays, specific promotional or marketing campaigns, logistics and production issues, and responses to anticipated increases or decreases in the costs of producing the product. Independent and dependent variables are both as defined in Chapter 3, when introducing regression.

For example, **Apple** might use the following independent variables in a regression analysis to predict future iPhone demand:

- Number of people who own previous models of the iPhone
- Economic factors (for example, growth or shrinkage of discretionary income)
- Perceived extent of upgrade over previous models
- Issues with production (such as logistics or production issues)
- Responses to anticipated cost increases or decreases

Based on your experience, which additional factors (independent variables) may predict iPhone demand? How will regression tell us if these factors help predict outcomes?

> **LAB CONNECTION**
>
> **LAB 5.6 EXCEL** provides step-by-step discussion of and practice with predicting product demand using independent variables (inputs) to predict the dependent (outcome) variable of product demand.

Other Examples of Regression in Predictive Analytics

We can use regression to answer a number of other important business questions, including the following.

How do the firm's costs depend on the level of production? Cost behavior analysis helps a company understand what portion of its production costs are fixed costs (costs that do not change with an increase or decrease in the quantity of goods or services produced) versus variable costs (costs that do change with an increase or decrease in the quantity of goods or services produced). Understanding these costs can help a company make better job-pricing and job-costing decisions.

Suppose that **Moisture Shield** is trying to understand the costs involved in the production of its wood/plastic composite decking. To understand these costs, Moisture Shield could run a regression, specifying the dependent (outcome) variable and the independent (input) variable to identify which production costs depend on the length of the boards (board feet) used to produce its decking.

- Dependent (outcome) variable: Total production costs
- Independent variable: Board feet (number of board feet of decking produced)

> **LAB CONNECTION**
>
> **LAB 8.4 EXCEL** (in Chapter 8) provides a step-by-step discussion of and practice with estimating cost behavior.

How do lenders set interest rates on loans they offer to borrowers? Based on various loan and borrower characteristics, business analysts can run a regression analysis to identify the factors that predict the interest rate that a lender charges a borrower.

- Dependent (outcome) variable: Interest rate charged to borrower
- Independent variables: Potential factors affecting the interest rate charged to the borrower, including:
  - Borrower's annual income
  - Level of debt compared to income (that is, the ratio of the borrower's total monthly debt payments to monthly income)
  - Credit history (including the number of times the borrower has been late with payments to creditors over the past two years)
  - Income verification status (income verified or not verified)
  - Rent/own (whether the borrower rents or owns their home)
  - Credit score (number assigned by credit agencies to suggest the borrower's ability to pay back a loan)

Time Series Analysis

Time series analysis predicts future values based on past values of the same variable. For example, if a company has a large amount of sales data over the current and past 12 or more quarters—that is, a *data series*—it can try to identify a trend that will help it predict future sales. **Persistence** is the repeatability (or continuity) and durability of critical variables such as earnings, sales, or cash flows, over time. Simply, it is an estimate of whether past trends will continue into the future.

Predicting Product Demand

As we saw above, a regression analysis takes into account the many independent variables that might predict product demand. In contrast, a time series analysis uses only one variable to predict the future. For example, if we are using time series analysis to predict future product demand, then the one variable that we use is past product demand, as shown in Exhibit 5.2.

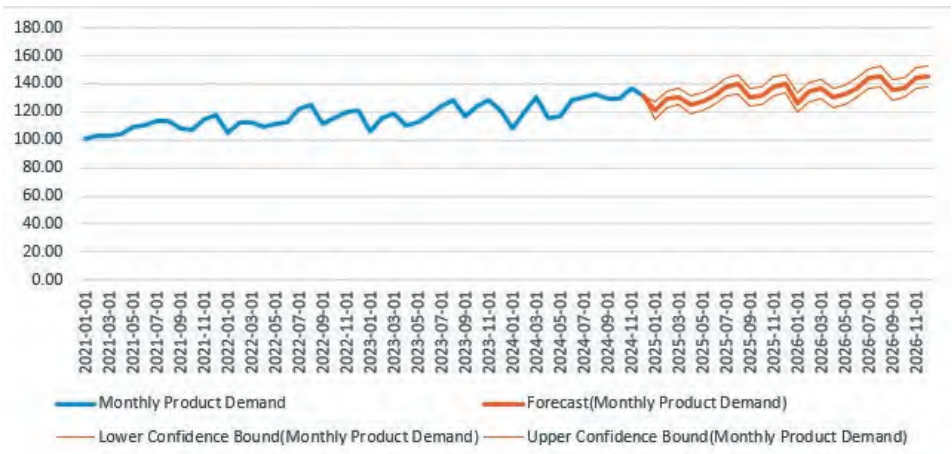


Exhibit 5.2
Time Series Analysis of Product Demand (from Lab 5.3 Excel)

In Exhibit 5.2, the past values of product demand from 2021 to 2024 (represented by the blue line) were used to predict product demand in 2025 and 2026 (represented by the bold orange line). The best-case scenario is the upper confidence bound and the worst-case scenario is the lower confidence bound, shown respectively in Exhibit 5.2 as the lighter orange lines above and below the bold orange line.

LAB CONNECTIONS

LABS 5.3 EXCEL and **LAB 5.4 TABLEAU** provide a step-by-step discussion of and practice with predicting product demand using earlier data regarding product demand. **LAB 5.6 EXCEL** uses a different technique to predict product demand.

PROGRESS CHECK

3. Why is predicting which firms we expect to go bankrupt considered predictive analysis?
4. Why is forecasting product demand using time series considered predictive analysis?

LO 5.3

Define *base rates* and *base rate fallacy*.

5.3 BASE RATES AND BASE RATE FALLACY

When performing classification and regression, it is sometimes important to conduct multiple levels of analysis. For example, a good first step in bankruptcy prediction is identifying how many companies go bankrupt each year (as a percentage of total firms). This information helps us understand the prevalence of bankruptcy as a whole. Here we are trying to identify the **base rate**, or the probability of an event occurring based on a related historical average.

Base rates are often used to gauge the reasonableness of a prediction. Suppose that our analysis suggests that 10 percent of companies will go bankrupt in the next two years, but the base rate of bankruptcies is only two percent. Perhaps our model overestimates the possibility of bankruptcy, or perhaps the economic conditions currently are much worse than in the past. A **base rate fallacy** occurs when the prediction places too little weight on the base rates and instead favors different or new information for the prediction. For example, imagine that a statistical model of the weather predicts rain on three percent of the days this summer in Northwest Arkansas due to El Niño, but the base rate in northwest Arkansas indicates rain 15 percent of the time during the summer. While the forecaster might believe there is good reason to ignore the base rate, most of the time the base rate should be carefully considered and incorporated into the analysis.

PROGRESS CHECK

5. Suppose that the marketing department believes that a new website will increase the conversion rate of website visitors into customers. When considering the possible consequences of a new website design, how important is knowing the base conversion rate? Explain.
6. Is it possible to put too much weight on the base rates? How do you find the middle ground between using long-term historical averages (base rates) and incorporating more current information?

5.4 PRESCRIPTIVE ANALYTICS

LO 5.4

Analyze data to determine the best course of action using prescriptive analytics and its techniques.

Prescriptive analytics tries to answer the questions "What should we do, based on what we expect will happen?" and "How do we optimize our performance based on potential constraints?"

Recall that predictive analytics helps us predict future outcomes. Based on those predicted outcomes, analysts and decision makers may adjust their decisions to achieve different and better outcomes. Prescriptive analytics performs what-if analysis, assessing whether changes to the inputs (and/or assumptions made) can lead to better outcomes.

Prescriptive analytics attempts to identify the best possible options given constraints or changing conditions, such as:

Constraints faced by the company. Every company has limited resources. Because firms don't have unlimited capital resources to invest, they cannot manufacture every product they would like and do not have an unlimited marketing budget. They must optimize given these and other constraints.

Changing conditions faced by the company. Most companies continuously face changing tastes in consumer preferences, changing tax laws, changing interest rates, changing levels of risk, changing exchange rates, and changing demand for employees and their expertise. These changing conditions require all companies to react and adjust to meet the needs of the market and environment. To the extent they can react and adjust better than their competition does, they will have a competitive advantage over other companies.

Let's now look at some specific examples of how prescriptive analytics can address a firm's constraints and the changing conditions that it faces.

Prescriptive Analytics Techniques

We will emphasize five prescriptive analytics techniques that are commonly used to address constraints and changing conditions:

1. What-if analysis, including marginal/incremental analysis
2. Cash flow analysis
3. Scenario analysis
4. Goal-seek analysis
5. Sensitivity analysis

Below, we discuss each type of analysis and how it relates to a specific business question.

Using What-If Analysis to Optimize Marketing Campaigns

What-if analysis is a prescriptive analytics technique that evaluates how potential changes in inputs may affect output. For example, what-if analyses can suggest how changes in the marketing mix (product, price, place, and/or promotion) will affect outcomes. For example, more social media advertising might bring in more revenue, or a new rewards program might positively impact customer loyalty and lead to higher sales.

Because companies do not have unlimited marketing resources, they must efficiently and effectively tailor their marketing campaigns to deliver the right message to the right people at the right time at the right price. To determine if its marketing campaign succeeded, the company will measure its impact using various impact metrics such as touchpoints, overall sales, and marketing exposure.

Marginal/incremental analysis is a type of what-if analysis that is typically used in economics, finance, or accounting settings. Management uses marginal/incremental analysis to determine the change in profit typically associated with the cost or benefit of the next (or marginal) unit. For example, a company may need to hire an additional employee to handle increased orders. However, before hiring a new employee, managers will want to know if the profits from those increased orders will pay the salary of the additional employee.

Using Cash Flow Analysis to Decide Whether to Pursue An MBA Degree

Suppose you are nearing the end of your senior year and you are asking yourself: Should I get an MBA now, or should I take a job and start my career now? To answer that question, you will want to know the decision's impact on your future income. Which option gives you the highest long-term payoff or the highest net present value (that is, value today)?

To answer your question, you can conduct a **cash flow analysis**, a prescriptive analytics technique that evaluates future cash flows for potential investments or expenditures, typically using net present value (NPV) or internal rate of return (IRR). Finance departments and investors perform similar cash flow analysis to assess the possible outcome of different investments. Chapters 8 and 9 provide examples of cash-flow analysis in the accounting and finance areas, along with a complete explanation of NPV and IRR.

LAB CONNECTION

LAB 5.7 EXCEL provides a step-by-step discussion of and practice with using cash flow analysis to assess the highest long-term payoff to enrolling in an MBA program.

Using Scenario Analysis to Determine the Expected Outcome of a Potential Trade War with Another Country

Let's assume there is a possibility of a trade war with another country. Three outcomes are possible: no effect (no trade war), a 10 percent decrease in company income, or a 25 percent decrease in company income. The CFO has asked you to evaluate how these three potential scenarios might affect the company. In other words, the CFO has asked you to perform a **scenario analysis**, a prescriptive analytics technique that predicts how changing multiple inputs at the same time will affect an outcome variable.

LAB CONNECTION

LAB 5.8 EXCEL provides a step-by-step discussion of and practice with using scenario analysis to assess the possible impact of a trade war. Specifically, we look at the impact on a business if import taxes are imposed and a trade war results.

Using Goal-Seek Analysis to Determine The Break-Even Level of Sales

Goal-seek analysis is a form of what-if analysis that tells us what needs to be done (or which inputs are needed) to achieve a desired outcome, output, or result. For example, companies often use goal-seek analysis to compute break-even sales, the level of sales at which revenues just cover expenses and result in a profit of $0.

Goal-seek analysis works backward, asking the analyst to identify the desired final outcome and then working backward to find the appropriate input(s). Calculating the break-even point is an essential part of understanding the point at which a company's sales exactly equal its variable and fixed costs, allowing the company to break even.

LAB CONNECTIONS

LAB 8.8 EXCEL (in Chapter 8) provides a step-by-step discussion of and practice with using goal-seek analysis to calculate the break-even level of sales.

Using Sensitivity Analysis to Determine the Impact of Changing Inputs on Outputs

Sensitivity analysis is what-if analysis that predicts how changing one input will affect an outcome variable. It is used in accounting and finance to understand how one specific input variable (such as interest rates, exchange rates, costs of capital, gas prices, or employee wages) might affect an outcome (such as expected sales, earnings, cash flows, or stock price). In other words, how sensitive are the expected, predicted outcomes to a range of possible input?

Exhibit 5.3 summarizes the prescriptive analysis techniques we've discussed, also providing example questions and the specific techniques employed. Future chapters provide more examples of prescriptive analytics, showing how we might use them to address changing conditions and constraints. For example, in Chapter 8 on operational analytics, we show how to use Excel's Solver Add-in to determine which transportation routes a company should use to ship product from multiple manufacturing plants to multiple warehouses.

Prescriptive Analysis Technique	Example Business Question	Specific Technique Employed
Marginal/incremental analysis: Determines the change in profit typically associated with the cost or benefit of the next (or marginal) unit	Should **Apple** sell its iPhones to the U.S. military for a lower price than it charges its other customers?	Evaluate the benefits (revenues) of selling to the U.S. military in terms of the cost of iPhone production.
Cash flow analysis Evaluates future cash flows for potential investments or expenditures, typically using net present value (NPV) or internal rate of return (IRR)	What cash flows are associated with investing in a new piece of equipment at **Boeing** or a new truck at **J.B. Hunt**?	Estimate the amounts, timing, and risk with respect to the predicted cash flows associated with investing in the new equipment.
Scenario analysis Predicts how changing multiple inputs at the same time will affect an outcome variable	How does **Wayfair** respond to various scenarios of increased taxes on goods imported from China?	Analyze all of the potential scenarios to arrive at the overall potential impact of the future events.
Goal-seek analysis Determines the required input value(s) to get a desired output (backward analysis)	For **McGraw Hill**, what level of sales is needed to break even on this textbook?	Perform a break-even analysis to figure out the needed level of sales. In terms of break-even sales, what is the level of input (sales) needed to get an output (profitability) of exactly 0?
Sensitivity analysis: Predicts how changing one input will affect an outcome variable	If the customer conversion rate from visits to **Amazon**'s website increases from five percent to seven percent or eight percent, what is the impact on the company's profits?	Test to see how sensitive Amazon's profits are to changing assumptions regarding conversion rates.

Exhibit 5.3 Prescriptive Analytics Techniques, Business Questions, and Specific Techniques Employed

DATA ANALYTICS AND YOUR CAREER

Accenture, a multinational professional services company that provides services in strategy, consulting, technology, and operations, suggests that accounting and finance professionals will need to have additional skills to handle technology and the mounds of data available. As summarized in Exhibit 5.4, finance and accounting professionals must anticipate alternative scenarios and answer "What should we do?" questions.

Exhibit 5.4
Analytics Skills Expectations for the Finance Workforce

Source: "Meet the 2020 Workforce," Accenture, https://www.accenture.com/t20170511t060010z__w__/us-en/_acnmedia/pdf-50/accenture-finance-2020-workforce.pdf. Accessed October 2021.

Digital innovations such as artificial intelligence, collaboration technologies and advanced analytics are rapidly disrupting the finance workforce. Traditional roles are evolving, and newer roles are increasing in importance.

Traditional roles		Main impacts
Financial planning and analysis		The real work starts when you deliver the reports or analysis: • Anticipating alternative scenarios, tracking their emergence and executing on contingency plans • Not just answering the "what happened" and "why did it happen" questions but also answering the "what should we do" questions
Financial controller		• Focusing on preventative and real-time control rather than relying on detective controls • Managing outcomes, not processes
Accounts payable clerk		• Focusing on exceptions as routine work is automated • Greater collaboration with other functions

Who is Danny/Shutterstock; fizkes/Shutterstock; EdBockStock/Shutterstock

PROGRESS CHECK

7. How might scenario analysis be applied to possible changes in tax rates due to a possible change in tax law? Why is scenario analysis considered to be part of prescriptive analytics?
8. What is sensitivity analysis, and why is it considered to be part of prescriptive analytics?

LO 5.5

Summarize the different types of data analytics and the business questions they attempt to answer.

5.5 SUMMARY OF ANALYTICS PERFORMED TO ADDRESS BUSINESS QUESTIONS

Exhibit 5.5 summarizes the five types of data analytics, the business questions that each attempts to answer, and the specific techniques that are employed as part of the analysis. Exhibit 5.6 uses a circumplex to summarize the analytics types and techniques we explained in Chapters 2–6. We provide more information on each analytics type and technique in Chapters 7–10 to bring together all of the analytics by business discipline (marketing, accounting, finance, and operations). In Chapter 11, we explain adaptive/autonomous analytics in more detail.

Data Analytics Type	Business Questions Analysts Try to Address	Definition of Analysis Type	Specific Techniques of Analysis Employed
Descriptive (Chapter 4)	What happened? What is happening?	Descriptive analytics characterizes, summarizes, and organizes features and properties of the data to facilitate understanding.	Counts, totals, sums, averages, financial statements, histograms, pivot tables, percentage change, ratio analysis, vertical and horizontal analysis

Exhibit 5.5
Data Analytics Types, Business Questions, Definitions. and Specific Techniques Employed

Exhibit 5.5 ***(Continued)***

Data Analytics Type	Business Questions Analysts Try to Address	Definition of Analysis Type	Specific Techniques of Analysis Employed
Diagnostic (Chapter 4)	Why did it happen? What are the reasons for past results? Can we explain why it happened?	Diagnostic analytics investigates the underlying cause that cannot be identified by simply looking at the descriptive data and may include hypothesis testing.	Hypothesis testing, variances, differences from expectations, correlations, visualize and cluster, identify outliers/anomalies, fuzzy lookup; drill-downs and roll-ups to get detail when needed; pivot tables (cross-tabulations), sequence check
Predictive (Chapter 5)	Will it happen in the future? What is the probability something will happen? Can we forecast what will happen?	Predictive analytics provides foresight by identifying patterns in historical data and assessing likelihood or probability.	Classifications, regressions, time series
Prescriptive (Chapter 5)	What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints?	Prescriptive analytics identifies best possible options given constraints or changing conditions.	What-if analysis, marginal/ incremental analysis, cash flow analysis, scenario analysis, goal-seek analysis, sensitivity analysis
Autonomous/ Adaptive (Chapter 11)	How can we continuously learn using artificial intelligence? Can we learn from past and current events with adaptive learning?	Adaptive/autonomous analytics is performed by machines that mimic human thought processes and self-learning to offer insight into large amounts of data.	Machine learning, simulations, Solver add-in, survival analysis S(), textual analysis

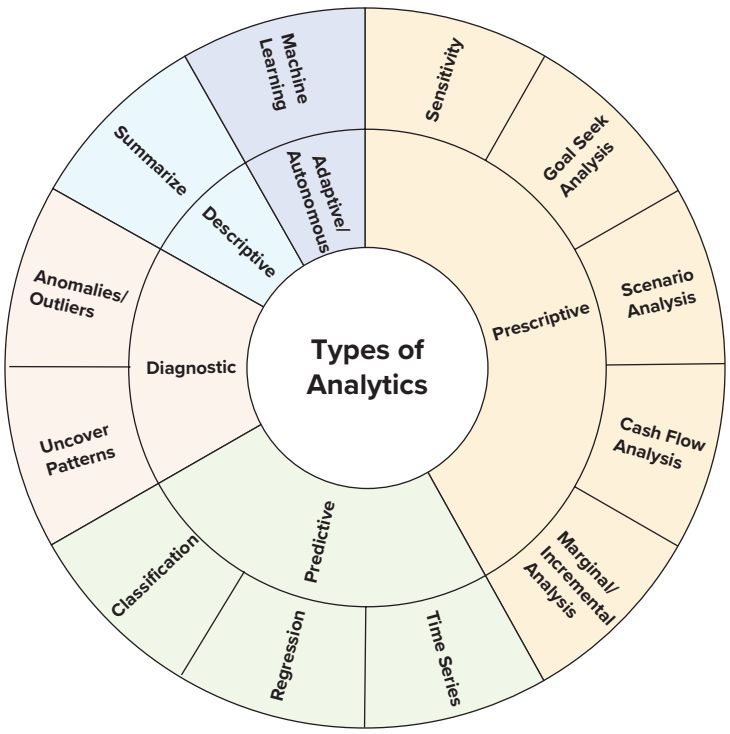


Exhibit 5.6
Analytics Types and Their Related Analytics Techniques

Summary

In this chapter, we explained that confirmatory data analytics emphasizes predictive and prescriptive analytics and their related techniques. Descriptive and diagnostic analytics are deterministic, but predictive and prescriptive analytics are probabilistic, judging the likelihood (probability) of a future event or outcome occurring.

We explained three common techniques in predictive analytics: classification, regression, and time series analysis. We also introduced five different techniques associated with prescriptive analytics: what-if analysis (including marginal/incremental analysis), cash-flow analysis, scenario analysis, goal-seek analysis, and sensitivity analysis.

Key Terms

base rate (*280*) Probability of some event occurring based on some historical average.

base rate fallacy (*280*) Occurs when a prediction places too little weight on the base rates and instead favors different or new information as the basis for the prediction.

cash flow analysis (*282*) A prescriptive analytics technique that evaluates future cash flows for potential investments or expenditures, typically using net present value (NPV) or internal rate of return (IRR).

classification (*274*) A predictive analytics technique to separate, segment, or classify a sample (or population) into two or more groups or classes.

confirmatory data analytics (*274*) Predictive and prescriptive analytics that use various techniques and statistics to evaluate the likelihood (probability) of a future event or outcome occurring.

credit score (*276*) A statistic that evaluates a customer's creditworthiness and ability to repay a loan.

customer purchase pattern (*276*) A means of classifying (or segmenting) customers according to what they have bought.

customer touchpoints (*275*) The ways in which a company engages with its customers.

fraud (*276*) Deception that is intended to result in financial gain.

goal-seek analysis (*282*) A form of what-if analysis that determines the required input value(s) to get a desired output (that is, backward analysis).

marginal/incremental analysis (*281*) Prescriptive analytics technique that determines the change in profit typically associated with the cost or benefit of the next (or the marginal) unit.

persistence (*279*) The repeatability (or continuity) and durability of critical variables, such as earnings, sales, or cash flows, over time; an estimate of whether past trends will continue into the future.

predictive analytics (*274*) Analytics performed to provide foresight by using statistical techniques to identify patterns in historical data and assess likelihood or probability.

prescriptive analytics (*281*) Analytics performed to identify the best possible options given constraints or expectations.

regression (*277*) A predictive analytics technique used to estimate an outcome based on variable inputs.

scenario analysis (*282*) A prescriptive analytics technique that predicts how changing multiple inputs at the same time will affect an outcome variable.

sensitivity analysis (*283*) A prescriptive analytics technique that predicts how changing one input will affect an outcome variable.

time series analysis (*279*) A predictive analytics technique used to predict future values based on past values of the same variable.

what-if analysis (*281*) A prescriptive analytics technique that evaluates how potential changes in inputs may affect output.

ANSWERS TO PROGRESS CHECKS

1. The balance in a bank account is known and is therefore deterministic. The chance that a company will go bankrupt is probabilistic because it is not known beforehand and is subject to chance.
2. Predicting rain on a certain day would be probabilistic because it is not known definitively which days it will rain or not rain in advance.
3. Predicting financial distress, such as the possibility of going bankrupt, is an important predictive analytics activity. Business analysts use classification to put companies into two classes or groups: those expected to go bankrupt and those *not* expected to go bankrupt in the near future.
4. Time series analysis is an important forecasting (and predictive analytics) technique. Using past values of product demand to predict future product demand is an example of predictive analytics.
5. When predicting the impact of a new website, marketing teams will want to know the company's base conversion rate, which is the rate at which current website visitors make purchases. This rate will help them understand and measure the change in the conversion rate after the launch of the new website.
6. It is possible to either overweight or underweight base rate information. Predictive analytics can be used to find the appropriate weight for both the long-term historical average (base rate) and more current information.
7. Recognizing the possible scenarios of tax-rate changes due to changing tax law is an example of prescriptive analytics. In this example, scenario analysis helps a company see the potential impact of different tax rates on its operations and profit.
8. Sensitivity analysis is a prescriptive analytics technique that predicts how changing one input affects an outcome variable.

connect

Multiple-Choice Questions

1. (LO 5.2) Which predictive analytics technique would a data analyst use to place a company's customers into different segments or categories?
   a. Classification
   b. Regression
   c. Time series analysis
   d. What-if analysis
2. (LO 5.2) Which predictive analytics technique would a data analyst use to predict product demand using past product demand values?
   a. Time series analysis
   b. Regression
   c. Classification
   d. Benford's Law

3. (LO 5.2) Which term refers to the repeatability (or continuity) and durability of critical variables such as earnings, sales, cash flows, and product demand?
   a. Time series analysis
   b. Persistence
   c. Base rate
   d. Volatility
4. (LO 5.3) Not considering historical averages when making predictions of possible future events is known as
   a. base rate fallacy.
   b. base rate ignorance.
   c. historical rate fallacy.
   d. historical rate ignorance.
5. (LO 5.4) Which type of analysis starts with the desired outcome and works backward to determine the necessary inputs?
   a. Scenario
   b. Sensitivity
   c. Goal-seek
   d. Cash flow
6. (LO 5.4) Predicting the impact of employee wages on net income is an example of which type of analysis?
   a. Scenario
   b. Sensitivity
   c. Goal-seek
   d. Cash flow
7. (LO 5.4) Determining a company's break-even level of sales is an example of which type of analysis?
   a. Scenario
   b. Sensitivity
   c. Goal-seek
   d. Cash flow
8. (LO 5.2) Assume that a data analyst uses regression to predict the interest rate that a bank will quote to a borrower. Which type of variable is the interest rate?
   a. Dependent
   b. Independent
   c. Random
   d. Predicted
9. (LO 5.2) Assume that we use weather as a factor to predict product demand. Which type of variable is weather?
   a. Dependent
   b. Independent
   c. Random
   d. Predicted
10. (LO 5.1) Unlike exploratory analytics, confirmatory analytics are
    a. probabilistic.
    b. deterministic.
    c. more challenging.
    d. less challenging.

11. (LO 5.2) Which scenario does an upper confidence bound in a time series represent?
    a. Most likely
    b. Worst-case
    c. Best-case
    d. Most unlikely

Discussion Questions

1. (LO 5.1) What is the difference between deterministic and probabilistic models? What is the difference between exploratory and confirmatory analytics?
2. (LO 5.2) How is predictive analytics different from descriptive and diagnostic analytics?
3. (LO 5.2) Why is a potential borrower's employment history an important component of a bank's determination whether to extend a loan to that borrower?
4. (LO 5.2) Why are credit history and credit score good predictors of the interest rate that a bank will charge a borrower?
5. (LO 5.2) Why is predicting what interest rate a lender might offer a regression exercise and not a classification activity?
6. (LO 5.2) Why is persistence important in using the past to predict the future?
7. (LO 5.2) In a time series analysis predicting future sales, what is the dependent (outcome) variable and what is the independent variable?
8. (LO 5.2) Why is persistence an important attribute for time series analysis?
9. (LO 5.3) Why do base rates need to be incorporated into a prediction activity?
10. (LO 5.3) Suppose that our analysis predicts that firms will go bankrupt 10 percent of the time, but base rates suggest that over the past 20 years 2 percent of firms have gone bankrupt. Is a base rate fallacy present here? Explain.
11. (LO 5.4) Why is a decision to sell additional units to a customer at a reduced price usually based on marginal analysis?
12. (LO 5.4) How would you set up the scenario analysis for a Chinese company with major exports to the United States if it faces a different level of potential tariffs (import taxes) from the U.S. government? How would the Chinese company consider the potential effect of 10, 25, or 50 percent tariffs?



Brief Exercises

1. (LO 5.2, 5.4, 5.5) Label each of the following activities as an example of descriptive, diagnostic, predictive, or prescriptive analytics.

Example of Analytics	**Type of Analytics** (Descriptive, Diagnostic, Predictive, or Prescriptive)
Bank reconciliation	
Fraud or no fraud expected in financial statements	
Cash flow analysis	
Estimating future stock market returns	
Counting the number of online customers	
Average cycle time of supply chain	

2. (LO 5.2, 5.4, 5.5) Label each of the following as the domain of descriptive, diagnostic, predictive, or prescriptive analytics.

Example of Analytics	**Type of Analytics** (Descriptive, Diagnostic, Predictive, or Prescriptive)
Benford's Law	
Loan acceptance/rejection by bank/lender	
Sequence check for missing checks	
Growth in customer base over the past five years	
Expected cash flow returns to an investment	

3. (LO 5.1, 5.2, 5.4) Label each of the following confirmatory analytics techniques as either predictive or prescriptive analytics.

Example of Analytics Technique Performed	**Predictive Analytics or Prescriptive Analytics?**
Analysis of whether to manufacture or outsource the manufacture of products	
Classification of customer segments	
Cash flow analysis for payoffs of capital investments	
Sensitivity of profits to changes in fuel costs	
Time series analysis of product demand	

Problems

1. (LO 5.2) Label each of the following examples of predictive analytics techniques as classification, regression, or time series analysis.

Example of Predictive Analytics	**Classification, Regression, or Time Series?**
Loan acceptance/loan rejection	
Interest rate given to borrowers	
Bankruptcy prediction	
Prediction of sales levels given research and development budgets	
Customer segmentation	
Using past product demand to forecast future product demand	
Fraud or no fraud expected in financial statements	
Using past company sales to forecast future company sales	

2. (LO 5.4) Identify which prescriptive analytics technique should be used to answer the following questions. Select from cash flow analysis, marginal/incremental analysis, goal-seek analysis, or scenario analysis.

Example of Prescriptive Analytics	Prescriptive Analytics Technique (Cash Flow Analysis, Marginal/Incremental Analysis, Goal-Seek Analysis, or Scenario Analysis)
Does it make financial sense to do one more printing of this textbook? Or is it better just to stick with the stock of textbooks now at the warehouse?	
What are the long-term cash flow consequences of an investment in new equipment?	
How will the number of our customers in Russia be affected by changing exchange rates between the U.S. dollar and the Russian ruble?	
What grade do I need to get on the final exam to get an "A" in this class?	
What is the overall impact on our customers of all possible outcomes of a trade war with a specific country?	
What level of sales is needed to break even with zero net profits?	
Should **Apple** make or outsource the manufacturing of its iPhones?	

3. (LO 5.2, 5.3) Suggest questions that **General Motors** might use for each of the following predictive analytics techniques to address. (*Hint:* Use Exhibits 5.1 and 5.3 as a guide.)
   a. Classification
   b. Regression
   c. Time series analysis
   d. Analysis of base rates
4. (LO 5.2) How might **Netflix** classify its customers into segments based on the following? Give examples of each and explain your answers.
   a. Communication touchpoints (websites, email, snail mail)
   b. Human touchpoints (interaction with a salesperson or customer service representative)
   c. Physical touchpoints (physical store location, product and packaging)
   d. Sensory touchpoints (fragrances, store lighting, store design)
5. (LO 5.4) Suggest operations analytics questions (that is, questions pertaining to the manufacturing and sale of products) that **Samsung** might address using each of the following prescriptive analytics techniques. (*Hint:* Use Exhibit 5.3 as a guide.)
   a. Sensitivity analysis
   b. Cash flow analysis
   c. Marginal/incremental analysis
   d. Goal-seek analysis
   e. Scenario analysis
6. (LO 5.1, 5.4) The cost of a car battery is a major component of the cost of an electric car. Which confirmatory analytics techniques (including prescriptive analytics and predictive analytics) should a business analyst use to answer the following questions?
   a. What is a battery manufacturer's break-even level of sales, based on the falling prices of car batteries?
   b. Should the cost of a car battery be a fixed cost or a variable cost at **Tesla**?
   c. How do profit margins change at differing levels of car battery prices at **Rivian**, an electric truck manufacturer (and a competitor to Tesla in the electric car space)?

Explain your answers.

7. (LO 5.2) If you are going to use regression to predict how many factory workers **Fox-Conn** needs to hire in order to manufacture the next model of the **Apple** iPhone, what is the dependent (outcome) variable, and what are several relevant independent variables? How would you determine if you have (1) selected useful and appropriate independent variables? (2) any missing variables?
8. (LO 5.4) Open the To College or Not to College data set in Excel to compare the salary levels (cash flows) of a college graduate. Your goal is to use the data to decide whether or not to get a college degree. In preparation for analyzing the data, answer the following questions.

Required

1. Evaluate the data set. What assumptions are being made regarding the cost of going to college and the subsequent salary levels?
2. What type of visualization would you use to compare the salaries of the two possibilities (college degree/no college degree)?
3. What additional data would be beneficial in addressing the question?
4. Which predictive and/or prescriptive analytics techniques should you use to gain insight into the question?



DATA VISUALIZATION

9. (LO 5.2) Open the Advertising Expenses and Sales data set in Excel. To complete the requirements in this problem, you will evaluate the association (negative vs. positive) between advertising expenditures and sales revenue. The data set contains the month, the date, the monthly advertising expenses, and revenues (sales).

Required



DATA VISUALIZATION

a. Create a scatterplot of the relationship between sales revenue and advertising expenses. Are the results consistent with your expectations? Does the relationship appear to be positive? Explain.
b. Use regression analysis to examine the relationship between advertising expenses and sales revenue, with sales revenues as the dependent (outcome) variable and advertising expenses as the independent variable. Is the association/relationship appear positive, or is it negative? For every $1 investment in advertising, what is the increase in revenues according to the regression?

Data Visualization Exercise

1. (LO 5.2) Open the Costs of Moistureshield Production data set in Excel. Analyzing these data will help **Moisture Shield**, a manufacturer of wood-plastic composite decking, determine its cost structure.

Required

a. Use a scatterplot to show the relationship between total production costs and total number of units produced. Add a trend line to visualize the relationship.
b. What are the variable and fixed costs according to the regression results? (The total costs are the dependent (outcome) variable. The independent variable is the number of units produced, called production).
c. Do the results for the variable costs appear to be reasonable?

CHAPTER 5 LABS

LAB 5.1 **Excel:** Evaluating the Relationship Between Sales and R&D Expenditures
LAB 5.2 **Excel:** Evaluating the Relationship Between Sales and R&D Expenditures: Testing for a Nonlinear Relationship
LAB 5.3 **Excel:** Forecasting Product Demand Using Time Series Analysis
LAB 5.4 **Tableau:** Forecasting Product Demand Using Time Series Analysis
LAB 5.5 **Power BI:** Forecasting Product Demand Using Time Series Analysis
LAB 5.6 **Excel:** Forecasting Product Demand Using Regression
LAB 5.7 **Excel:** Assessing the Returns to Education: To MBA or Not to MBA?
LAB 5.8 **Excel:** Applying Scenario Analysis: Possible Trade War

Lab 5.1 Excel

Evaluating the Relationship Between Sales and R&D Expenditures

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Research and Development Expenditures, Sales Revenue, Regression

Decision-Making Context

Research and development (R&D) expenditures are investments that companies make to create new insights and new products that customers might like. Over time, these investments may pay off in the form of increased revenues and profits. This lab examines the possible payoff to R&D expenditures in the form of changed revenues.

Companies can run a regression to understand the relationship between sales revenue and R&D expenditures. The outcome variable is sales revenue, which arguably results from changes in the input variable, R&D expenditures.

- Dependent (outcome) variable: Sales revenue
- Independent variable: R&D expenditures

Required

Using regression analysis, determine the relationship between R&D expenditures and sales revenue.

Specify the Question

Does regression analysis explain the relationship between R&D expenditures and sales revenue?

Obtain the Data

Open Excel file Lab 5.1 Data.xlsx and browse around a bit.

Data Dictionary

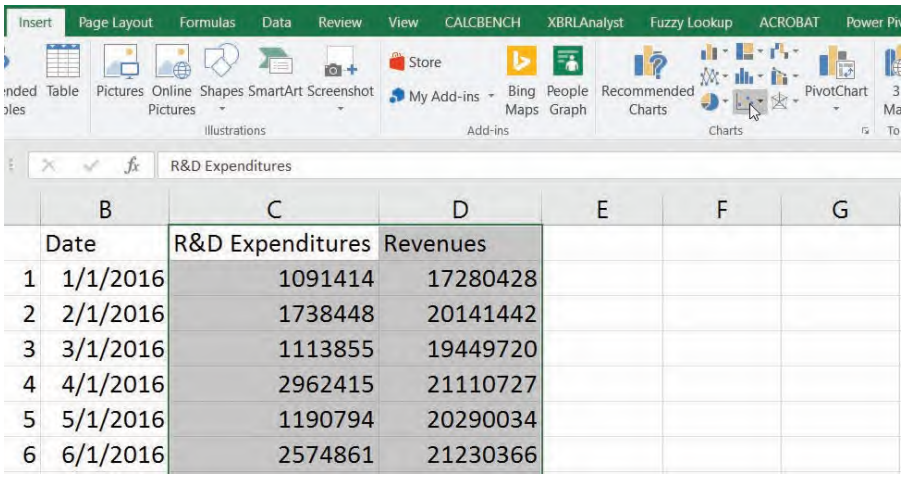
Month:	The month number from the beginning of the analysis until the end of the analysis.
Date:	Month when revenues and R&D expenditures are realized.
R&D Expenditures:	The monthly R&D expenditures for this company.
Revenues:	The monthly revenue for this company.

Analyze the Data

Analysis Task 1: Graph the Relationship Between R&D Expenditures and Sales.

1. Let's start by plotting the data. Highlight cells C1:D109. Select **Insert > Charts > Scatter** from the menu bar (Lab Exhibit 5.1.1).

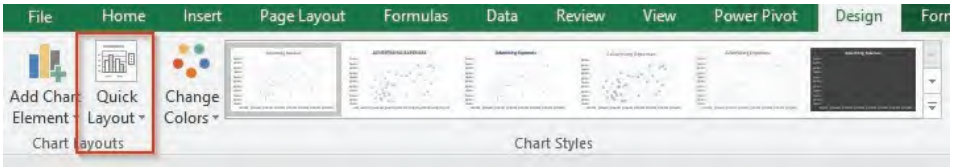
Lab Exhibit 5.1.1



	B	C	D	E	F	G
	Date	R&D Expenditures	Revenues			
1	1/1/2016	1091414	17280428			
2	2/1/2016	1738448	20141442			
3	3/1/2016	1113855	19449720			
4	4/1/2016	2962415	21110727			
5	5/1/2016	1190794	20290034			
6	6/1/2016	2574861	21230366			

2. Choose Layout 3 from Quick Layout on the ribbon (Lab Exhibit 5.1.2).

Lab Exhibit 5.1.2

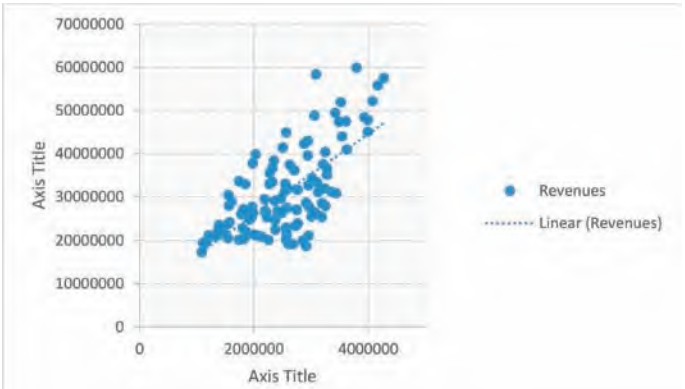




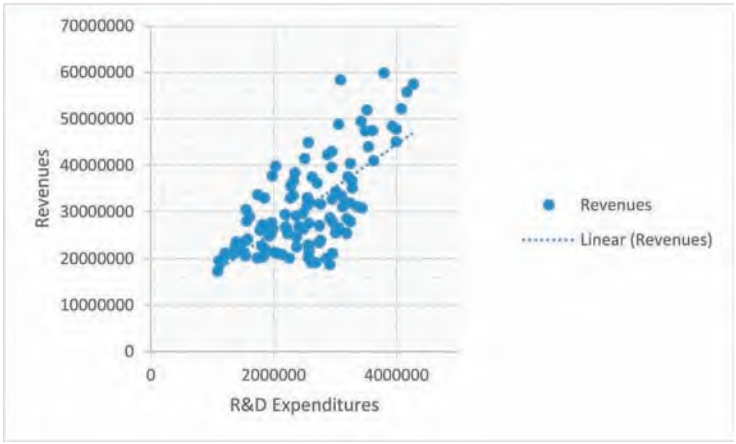
DATA VISUALIZATION

The result is a scatterplot plotting revenues and expenses, as shown in Lab Exhibit 5.1.3.

Lab Exhibit 5.1.3



3. Label the X-axis "R&D Expenditures" and the Y-axis "Revenues", as shown in Lab Exhibit 5.1.4.



Lab Exhibit 5.1.4

Given this graph, does it appear that there is a relationship between R&D expenditures and revenues? If so, does the relationship appear to be positive?
R&D expenditures are on the X-axis because we believe that R&D expenditures might be a predictor of revenues. In other words, revenue as a function of this type of expenditures = f(R&D expenditure).

Analysis Task 2: Perform Regression Analysis.

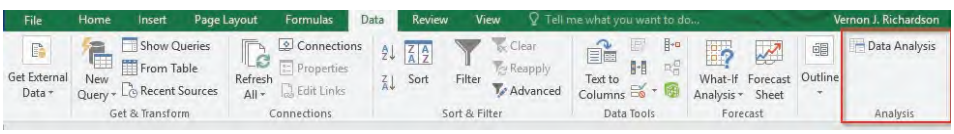
How can we determine if there is a statistical relationship between R&D expenditures and sales revenue? To consider this relationship, we will run a regression specifying the following:

- Dependent (outcome) variable: Sales revenue
- Independent variable: R&D expenditures

Regression analysis uses mathematics and statistics to determine which of several variables affect an outcome and the size of that impact. Regression analysis helps to answer the questions "Which factors matter most in predicting the outcome?" and "Which factors can we ignore?"

Before performing the regression, it is a good idea to think through the variables. For example, it seems likely that higher R&D expenditures are associated with higher sales. Therefore, we expect a positive (+) relationship between R&D expenditures and revenues.

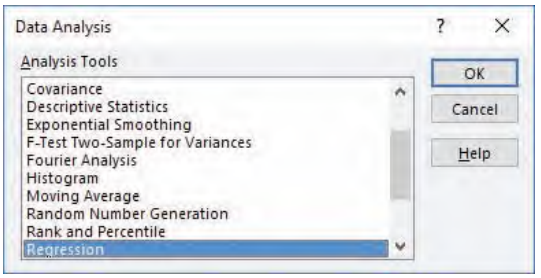
1. We need the Data Analysis ToolPak to run the regression. If you haven't installed the Data Analysis ToolPak, Appendix A provides instructions for doing so.
2. To run the regression, make sure the Data Analysis ToolPak is loaded by navigating to **Data > Analysis** in the ribbon and seeing if the Data Analysis Add-In is installed (Lab Exhibit 5.1.5).



Lab Exhibit 5.1.5

3. To perform the regression, select **Data > Analysis > Data Analysis.** A dialog box will open. Select regression, as shown in Lab Exhibit 5.1.6.

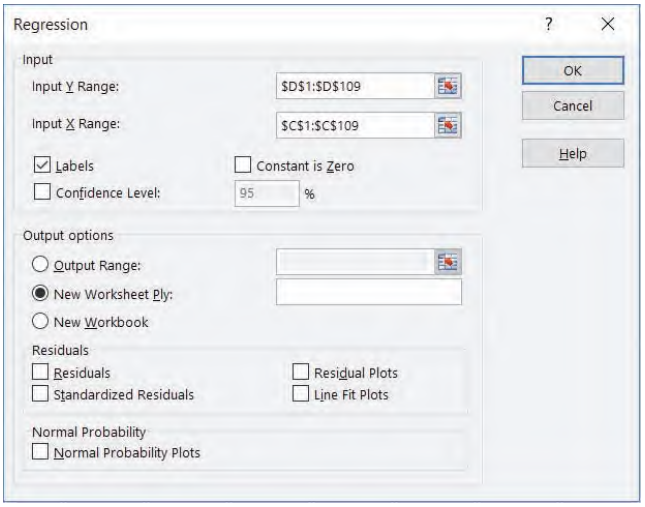
Lab Exhibit 5.1.6



4. A dialog box will open. We need to select the dependent (outcome) variable (or Y Range). Our dependent (outcome) variable is Revenues, so enter $D$1:$D$109 in the Input Y Range.
5. Next we need to select our independent variable (or X Range). The variable R&D expenditures is our independent variable, so enter $C$1:$C$109 in the Input X Range.
6. Because we highlighted the labels in the data range, click on "Labels". Then select "New Worksheet Ply" to get the results in a new worksheet.

The completed Regression dialog box should look as shown in Lab Exhibit 5.1.7.

Lab Exhibit 5.1.7



7. Click OK to get the regression results, which are shown in Lab Exhibit 5.1.8.

Lab Exhibit 5.1.8

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.686491894				
5	R Square	0.471271121				
6	Adjusted R Square	0.466283113				
7	Standard Error	7400410.253				
8	Observations	108				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	1	5.17434E+15	5.17434E+15	94.48082	2.39098E-16
13	Residual	106	5.8052E+15	5.47661E+13		
14	Total	107	1.09795E+16			
15						
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
17	Intercept	7387039.008	2527280.247	2.922920406	0.00424	2376460.11
18	R&D Expenditures	9.281064367	0.954829777	9.720124558	2.39E-16	7.388021465
19						

There are several things to note about the regression. First, the regression performed better at predicting revenues than a prediction made by random chance. We learn this by looking at the "Significance F" result in Row 12, which is very small, suggesting there is virtually zero probability that revenues can be explained by no independent variables. This result suggests that we will be able to identify factors that explain revenues.

We can use another statistic to measure how R&D expenditures affect revenues. The Adjusted R Square in Row 6 is a value between 0 and 1. *Note:* Statisticians prefer the term "adjusted R-squared," but Excel doesn't use that term. An adjusted R-squared value of 0 indicates that the model has no ability to explain the dependent (outcome) variable, and an adjusted R-squared value of 1 indicates that the model has perfect ability to explain the dependent (outcome) variable. In this case, the adjusted R-squared value is 0.466, which represents a moderate ability to explain the level of revenues given R&D expenditures.

Finally, the t Stat in Column D indicates whether R&D expenditures help predict revenues. The t Stat is greater than two for the independent variable, R&D Expenditures, suggesting that the greater the amount of R&D expenditures, the greater the total sales revenue.

Report the Results

Using scatterplots and regression analysis, we find evidence that the relationship between R&D expenditures and revenues is positive. At this point, management needs to decide if the returns to the R&D expenditures are worth their expected payoff in terms of sales. Managers could also estimate the increased income the company might receive based on higher R&D expenditures to determine if R&D is worthwhile.

Assessment

1. Take a screenshot of the scatterplot of the relationship between R&D expenditures and revenues, paste it into a Word document named "Lab 5.1 Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the regression output predicting revenues, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 5.1 Submission.docx via Connect, if assigned.

Alternate Lab 5.1 Excel

Evaluating the Relationship Between Sales and R&D Expenditures

Apply the same steps in Lab 5.1 to the Alt Lab 5.1 Data.xlsx data set.

Required

Using regression analysis, determine the relationship between R&D expenditures and sales revenue.

Open Excel file Alt Lab 5.1 Data.xlsx.

Data Dictionary

Month:	The month number from the beginning of the analysis until the end of the analysis.
Date:	Month when revenues and R&D expenditures are realized.
R&D Expenditures:	The monthly R&D expenditures for this company.
Revenues:	The monthly revenue for this company.

1. Take a screenshot of the scatterplot of the relationship between R&D expenditures and revenues, paste it into a Word document named “Alt Lab 5.1 Submission.docx”, and label the screenshot Submission 1.
2. Take a screenshot of the regression output predicting revenues, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 5.1 Submission.docx via Connect, if assigned.

Lab 5.2 Excel

Evaluating the Relationship Between Sales and R&D Expenditures: Testing for a Nonlinear Relationship

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Research and Development (R&D) Expenditures, Sales Revenue, Regression

Decision-Making Context

Research and development (R&D) expenditures are investments that companies make to create new insights and new products that customers might like. Over time, these investments may pay off in the form of increased revenues and profits.

In Lab 5.1, you determined whether there is a linear relationship between revenues and R&D expenditures. This lab goes a step further to determine whether the relationship is nonlinear. We test for nonlinearity by adding an additional term, R&D expenditures squared, or $(\text{R\&D expenditures})^2$.

Companies can run a regression to understand the relationship between sales revenue and R&D expenditures. The outcome variable is sales revenue, which arguably results from changes in the input variable, R&D expenditures.

- Dependent (outcome) variable: Sales revenue
- Independent variable: R&D expenditures, $(\text{R\&D expenditures})^2$

Required

Using regression analysis, determine the relationship between R&D expenditures, R&D expenditures squared, and sales revenue.

Specify the Question

Is there a nonlinear relationship between R&D expenditures and sales revenue?

Obtain the Data

Open Excel file Lab 5.2 Data.xlsx and browse around a bit to familiarize yourself with the data.

Data Dictionary

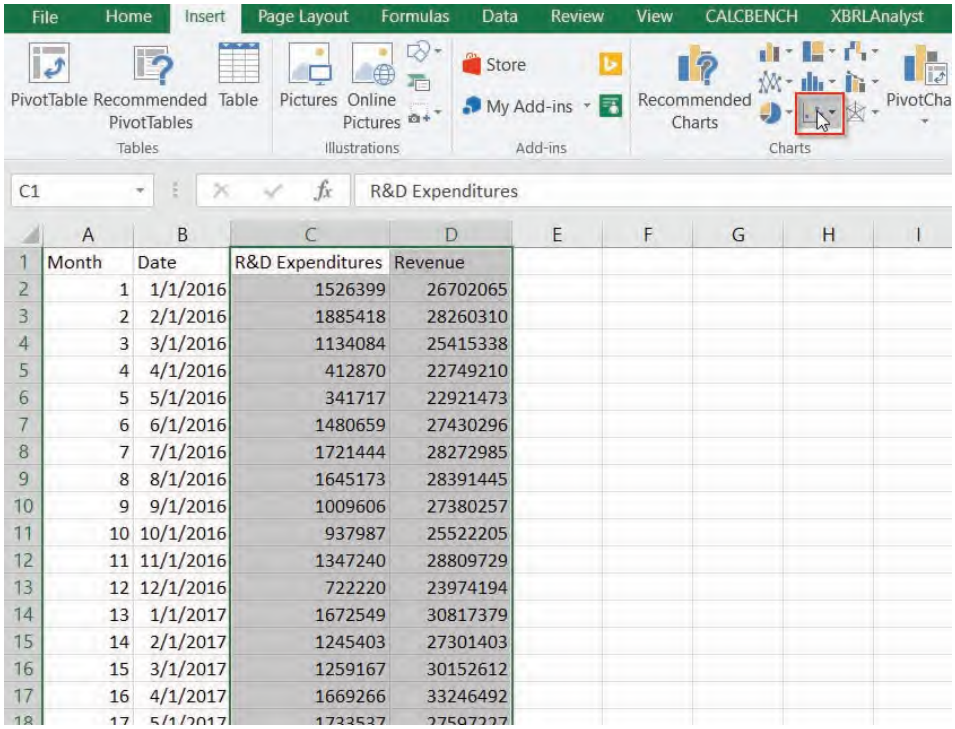
Month:	The month number from the beginning of the analysis until the end of the analysis.
Date:	Month when revenues and R&D expenditures are realized.
R&D Expenditures:	The monthly R&D expenditures for this company.
Revenue:	The monthly revenue for this company.

Analyze the Data

Analysis Task 1: Graph the Relationship Between R&D Expenditures and Sales.

1. Let's start by plotting the data. Highlight cells C1:D109. Select **Insert > Charts > Scatter** from the menu bar (Lab Exhibit 5.2.1).

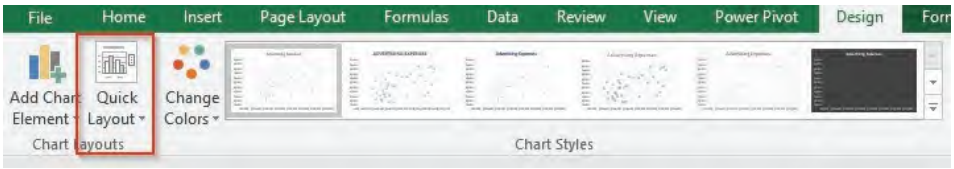
Lab Exhibit 5.2.1



	A	B	C	D	E	F	G	H	I
1	Month	Date	R&D Expenditures	Revenue					
2	1	1/1/2016	1526399	26702065					
3	2	2/1/2016	1885418	28260310					
4	3	3/1/2016	1134084	25415338					
5	4	4/1/2016	412870	22749210					
6	5	5/1/2016	341717	22921473					
7	6	6/1/2016	1480659	27430296					
8	7	7/1/2016	1721444	28272985					
9	8	8/1/2016	1645173	28391445					
10	9	9/1/2016	1009606	27380257					
11	10	10/1/2016	937987	25522205					
12	11	11/1/2016	1347240	28809729					
13	12	12/1/2016	722220	23974194					
14	13	1/1/2017	1672549	30817379					
15	14	2/1/2017	1245403	27301403					
16	15	3/1/2017	1259167	30152612					
17	16	4/1/2017	1669266	33246492					

2. Choose Layout 3 from the Quick Layout on the ribbon, as shown in Lab Exhibit 5.2.2.

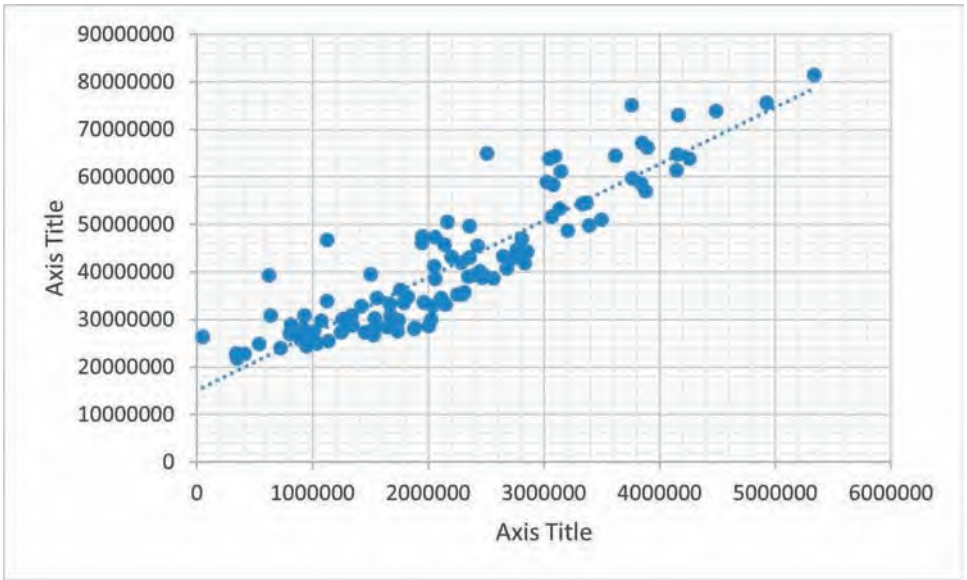
Lab Exhibit 5.2.2



DATA VISUALIZATION

The result is a scatterplot plotting revenues and expenses, as shown in Lab Exhibit 5.2.3.

Lab Exhibit 5.2.3

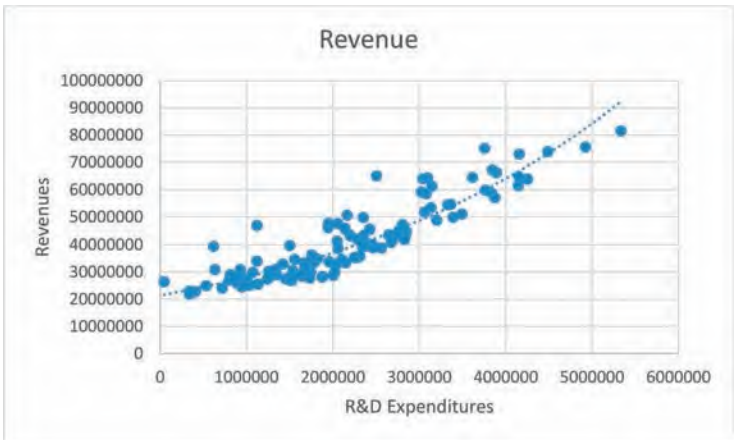


3. Next, select **Add Chart Element > Trendline > More Trendline Options > Exponential,** as shown in Lab Exhibit 5.2.4.



Lab Exhibit 5.2.4

4. Label the X-axis as "R&D Expenditures" and the Y-axis as "Revenues", as shown in Lab Exhibit 5.2.5.



Lab Exhibit 5.2.5

Given this graph, does it appear that there is a relationship between R&D expenditures and revenues? Does this relationship look nonlinear?

Does the relationship appear to be positive? That is, as R&D expenditures increase, do revenues also increase?

We put the R&D expenditures on the X-axis because we believe that R&D expenditures might be a predictor of revenues. In other words, revenues are a function of these type of expenditures = f(R&D expenditure).

Analysis Task 2: Perform Regression Analysis.

How can we determine if there is a statistical relationship between R&D expenditures and sales revenue? To consider this relationship, we will run a regression specifying the following:

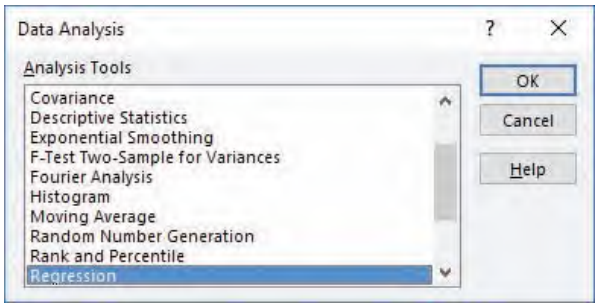
- Dependent (outcome) variable: Sales revenue
- Independent variables: R&D expenditures, $(\text{R\&D expenditures})^2$

1. We need to insert and calculate $(\text{R\&D expenditures})^2$. To do so, highlight the full column D by clicking on column D at the top. Right-click and select Insert. Next, in cell D1, type "R&D Expenditures Squared".
2. In cell D2, insert the formula "=C2^2" and hit Enter. Copy the result from cell D2 to cells D3:D109.

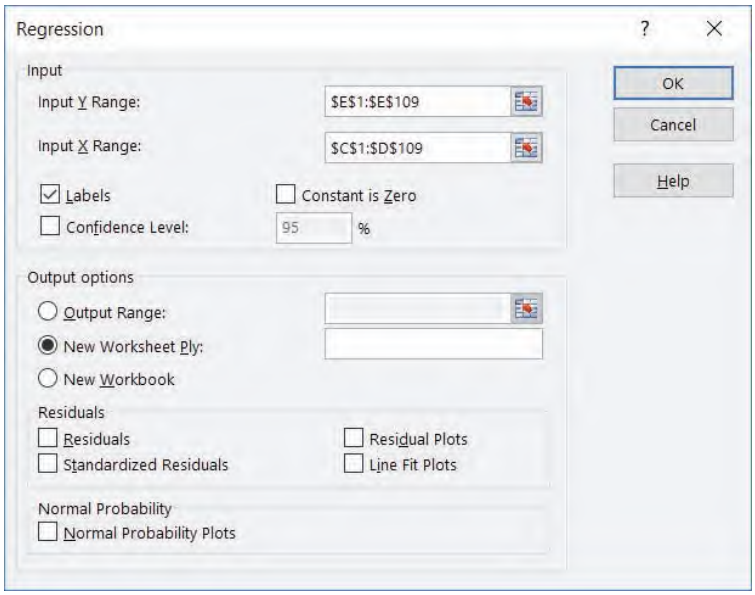
   Regression analysis uses mathematics and statistics to determine which of several variables affect an outcome and the size of that impact. Regression analysis helps to answer the questions "Which factors matter most in predicting the outcome?" and "Which factors can we ignore?" Before performing the regression, it is a good idea to think through the variables. For example, it seems likely that higher R&D expenditures are associated with higher sales. Therefore, we expect a positive (+) relationship between R&D expenditures and revenues, but at an increasing rate as shown by the nonlinear exponential trend line determined in Step 1.

   We're now ready to perform the regression analysis.
3. To perform the regression, select **Data > Analysis > Data Analysis.** A dialog box will open. Select Regression, as shown in Lab Exhibit 5.2.6.

Lab Exhibit 5.2.6



4. A dialog box will open. We need to select the dependent (outcome) variable (or Y Range). Our dependent (outcome) variable is revenues, so enter " $E$1:$E$109" in the Input Y Range.
5. Next we need to select our independent variable (or X Range). The variables R&D expenditures and $(\text{R\&D Expenditures})^2$ are our independent variables, so enter "$C$1:$D$109" in the Input X Range.
6. Because we highlighted the labels in the data range, click on Labels. Also select New Worksheet Ply to get the results in a new worksheet.
7. The completed Regression dialog box should look as shown in Lab Exhibit 5.2.7.



Lab Exhibit 5.2.7

8. Click OK to get the regression results, which are shown in Lab Exhibit 5.2.8.

Lab Exhibit 5.2.8

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.909478				
5	R Square	0.827151				
6	Adjusted R Square	0.823859				
7	Standard Error	6030121				
8	Observations	108				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	2	1.82709E+16	9.14E+15	251.2333	9.49532E-41
13	Residual	105	3.81805E+15	3.64E+13		
14	Total	107	2.20889E+16			
15						
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
17	Intercept	21367225	2269617.303	9.414462	1.26E-15	16866993.19
18	R&D Expenditures	5.501598	1.980307943	2.778153	0.006477	1.575013356
19	R&D Expenditures Squared	1.31E-06	3.89595E-07	3.351151	0.001119	5.33096E-07

There are several things to note about the regression. First, the regression performed better at predicting revenues than a prediction made by random chance. We learn this by looking at the "Significance F" result in Row 12, which is very small, suggesting there is virtually zero probability that revenues can be explained by no independent variables. This result suggests that we will be able to identify the factors that explain revenues.

We can use another statistic to measure how well R&D expenditures predict revenues. The Adjusted R Square in Row 6 is a value between 0 and 1. *Note:* Statisticians prefer the term "adjusted R-squared," but Excel doesn't use that term. An adjusted R-squared value of 0 indicates that the model has no ability to explain the dependent (outcome) variable, and an adjusted R-squared value of 1 indicates that the model has perfect ability to explain

the dependent (outcome) variable. In this case, the adjusted R-squared value is 0.824, which represents a high ability to explain the level of revenues given R&D expenditures.

Finally, the t Stat in Column D indicates whether R&D expenditures help predict revenues. The t-Stat is greater than 2 for the independent variables, R&D Expenditures and R&D Expenditures squared, suggesting that the greater the amount of R&D expenditure, the greater the total sales revenue. And, given the positive (although very small) coefficient on R&D Expenditures Squared, revenue seems to be increasing at a relatively small increasing rate.

Report the Results

We have now found the relationship between R&D expenditures (Lab 5.1), R&D expenditures squared, and revenues, and we have found the relationship to be positive. At this point, management needs to decide if the R&D expenditures are worth their expected payoff in terms of predicted sales.

Assessment

1. Take a screenshot of the scatterplot of the relationship between R&D expenditures and revenues, paste it into a Word document named "Lab 5.2 Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the regression output predicting revenues, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 5.2 Submission.docx via Connect, if assigned.

Alternate Lab 5.2 Excel

Evaluating the Relationship Between Sales and R&D Expenditures: Testing for a Nonlinear Relationship

Apply the same steps in Lab 5.2 to the Alt Lab 5.2 Data.xlsx data set.

Required

Using regression analysis, determine the relationship between R&D expenditures squared and sales revenue.

Open Excel file Alt Lab 5.2 Data.xlsx.

Data Dictionary

Month:	The month number from the beginning of the analysis until the end of the analysis.
Date:	Month when revenues and R&D expenditures are realized.
R&D Expenditures:	The monthly R&D expenditures for this company.
Revenues:	The monthly revenue for this company.

Required

1. Take a screenshot of the scatterplot of the relationship between R&D expenditures and revenues, paste it into a Word document named "Alt Lab 5.2 Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the regression output predicting revenues, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 5.2 Submission.docx via Connect, if assigned.

Lab 5.3 Excel

Forecasting Product Demand Using Time Series Analysis

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Predictive Analysis, Forecasting, Time Series, Product Demand

Decision-Making Context

There are many reasons to forecast future product demand. For example, future product demand is used to ensure that a company has adequate product available for sale.

Hike 'Em Hiking Boots is a fictitious company that sells hiking boots and works hard to have the right products on hand and available for sale.

Required

Using Excel's forecasting sheet and available product demand from 2021 to 2024, forecast product demand for Hike 'Em Hiking Boots from 2025 to 2026.

Specify the Question

What will be the monthly demand for Hike 'Em Hiking Boots during 2025 and 2026?

Obtain the Data

Open Excel file Lab 5.3 Data.xlsx.

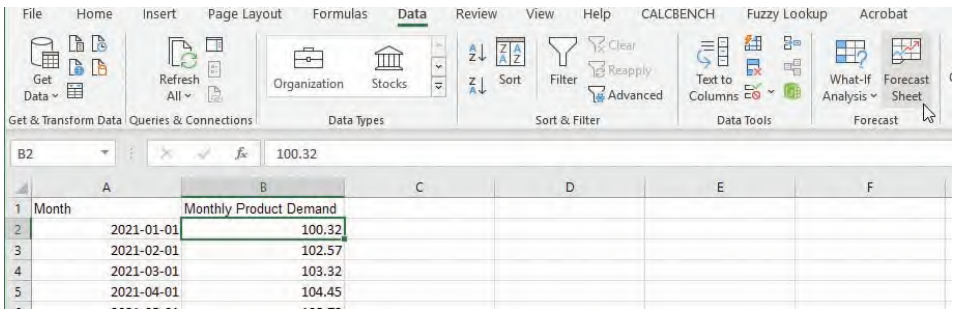
Data Dictionary

Month:	Month of the year from 2021 to 2024
Monthly Product Demand:	Product demand (in thousands of dollars sold)

Analyze the Data

Analysis Task 1: Analyze Past Trends.

1. To start the forecast analysis, highlight one of the numbers in column B (Monthly Product Demand), and then select **Data > Forecast > Forecast Sheet**, as shown in Lab Exhibit 5.3.1.



Lab Exhibit 5.3.1

The dialog box in Lab Exhibit 5.3.2 will open, beginning our forecast of Hike 'Em Hiking Boots sales for 2025 and 2026. The forecast sheet capability is available only on Excel for Windows, and not on Excel for Mac. If you work mostly on a Mac, you may need to use a PC in a computer lab or a virtual desktop to access this functionality.

Lab Exhibit 5.3.2



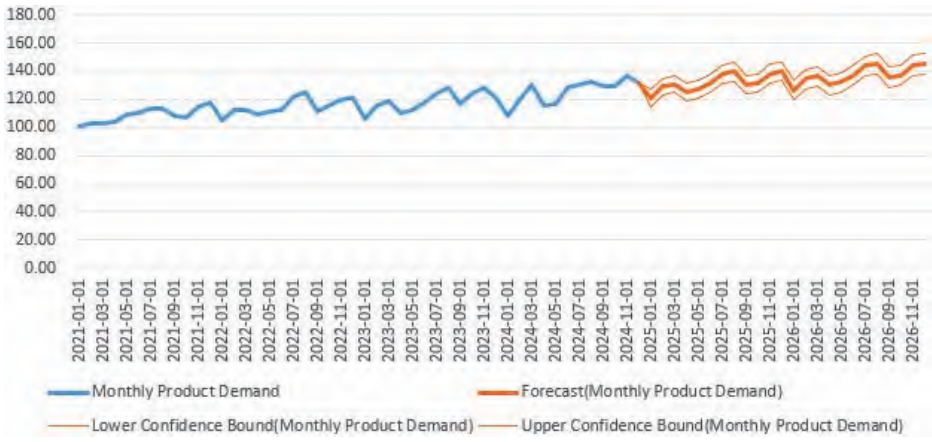
2. Note the seasonality of the data with the reasonably consistent crest and trough over the four quarters each year. Click on the chevron next to Options near the bottom left.
3. Consider the various options, including the dates that the forecast starts and ends, the confidence interval, the timeline range and value range, and how missing points and duplicates are handled.

Analysis Task 2: Forecast Future Performance.

1. Select Include forecast statistics near the bottom left and then select Create.
2. Note the new spreadsheet tab with forecast statistics and forecasts for each month in 2025 and 2026. Also note the lower and the upper confidence bound of the forecast. The forecast sheet suggests that there is a 95% chance that the actual sales will be in the range between the lower and the upper confidence bound.
3. The forecast graph should look as shown in Lab Exhibit 5.3.3.

DATA VISUALIZATION

Lab Exhibit 5.3.3



Report the Results

We have now developed a forecast for future product demand that the management of Hike 'Em Hiking Boots can use to plan its future operations and sales. In addition to using time series analysis, we can work to develop other predictors of product demand beyond prior product demand. Perhaps we can determine if there are other relevant predictors of performance, such as competitor performance, industry performance, or macroeconomic performance.

Assessment

1. Take a screenshot of the forecast spreadsheet of product demand, paste it into a Word document named "Lab 5.3 Excel Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the forecast graph of product demand, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 5.3 Excel Submission.docx via Connect, if assigned.

Alternate Lab 5.3 Excel

Forecasting Product Demand Using Time Series Analysis

Using Excel's forecasting sheet and available product demand, forecast product demand for Adventure Shoes from 2025 to 2026. Open Excel file Alt Lab 5.3 Data.xlsx.

Data Dictionary

Month:	Month of the year from 2021 to 2024
Monthly Product Demand:	Product demand (in thousands of dollars sold)

Assessment

1. Take a screenshot of the forecast graph of product demand, paste it into a Word document named "Alt Lab 5.3 Excel Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the forecast spreadsheet of product demand, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 5.3 Excel Submission.docx via Connect, if assigned.

Lab 5.4 Tableau

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Forecasting Product Demand Using Time Series Analysis

Keywords

Predictive Analysis, Forecasting, Time Series, Product Demand

Decision-Making Context

There are many reasons to forecast a firm's future performance. For example, predicting sales dollars may help a company determine its needed manufacturing capacity. Predicting cash flows may help the company decide if it will need a loan or not. At the individual level, an investor might forecast a company's earnings to evaluate a potential investment.

The goal of this lab is to forecast sales and earnings for Hike 'Em Hiking Boots, a fictitious company.

Required

Using Tableau's forecasting options, forecast product demand for hiking boots at Hike 'Em Hiking Boots from 2025 to 2026.

Specify the Question

What will be the monthly demand (in boot sales) for Hike 'Em Hiking Boots during 2025 through 2027?

Obtain the Data

Open Tableau and connect to Excel file Lab 5.4 Data.xlsx.

Data Dictionary

Month:	Month of the year from 2021 to 2024
Monthly Product Demand:	Product demand (in dollar sales of boots)

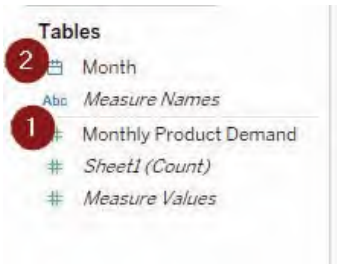
Perform the Analysis

Analysis Task 1: Plot the Time Series Data.

1. To start the forecast analysis, click on Sheet 1, double-click **Monthly Product Demand** from Measure Names, then double-click **Month** from Tables (Lab Exhibit 5.4.1).

Lab Exhibit 5.4.1

Tableau Software



2. Tableau should have defaulted the visualization type to a continuous line chart, but if it did not, you can change it from the **Show Me** tab (Lab Exhibit 5.4.2).



Lab Exhibit 5.4.2

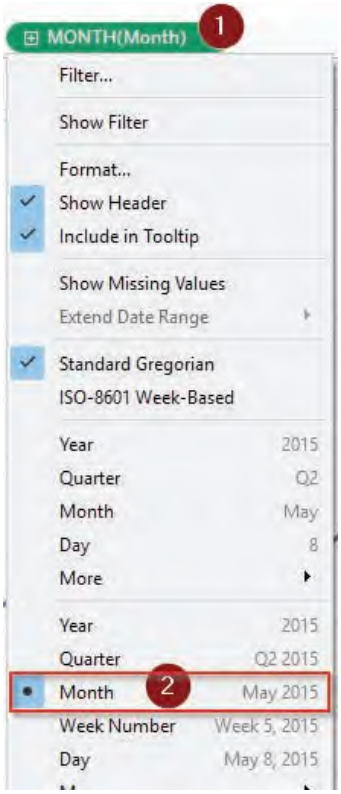
Tableau Software

Tableau interacts with dates in two ways, as either "Continuous" or "Discrete." Regardless of the way Tableau chose to display your dates when you added them to your visualization, you can change the field to the right display quite easily.

3. From the Columns shelf, right-click the YEAR(month) pill, and select Month May 2015 (Lab Exhibit 5.4.3).

Lab Exhibit 5.4.3

Tableau Software

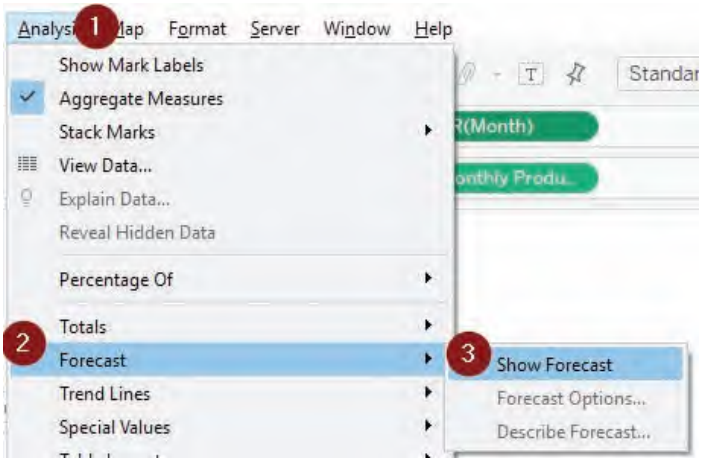


Analysis Task 2: Forecast Future Performance.

4. To expand this line chart to show a Forecast, select **Analysis** Tab > **Forecast** > **Show Forecast** (Lab Exhibit 5.4.4).

Lab Exhibit 5.4.4

Tableau Software



You can now select from a variety of options to refine this Forecast.

5. From the Marks shelf, change the default mark to Automatic and Circle (as shown in Lab Exhibit 5.4.5).

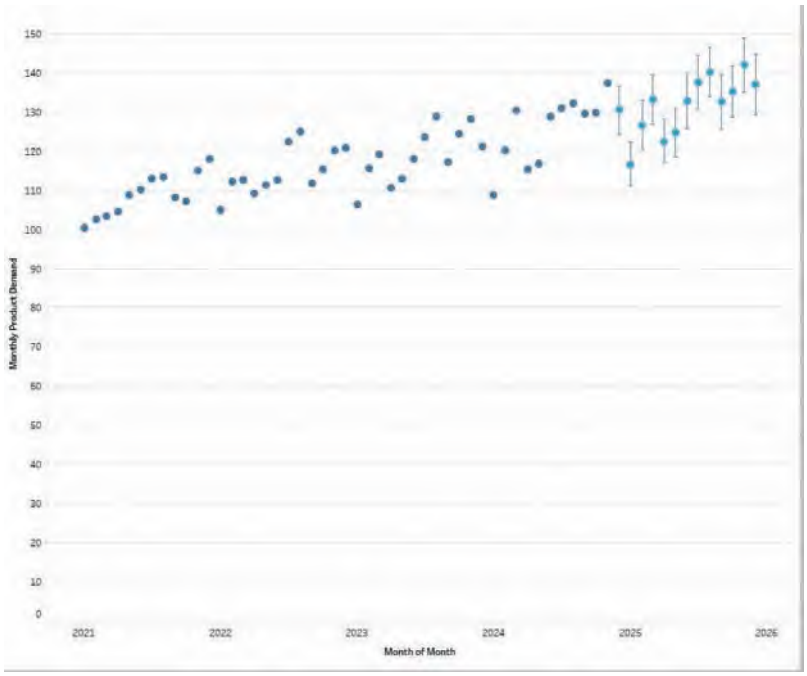


Lab Exhibit 5.4.5

Tableau Software

The line chart has changed, and Tableau has provided whiskers around each of the forecasted data points (Lab Exhibit 5.4.6).

DATA VISUALIZATION



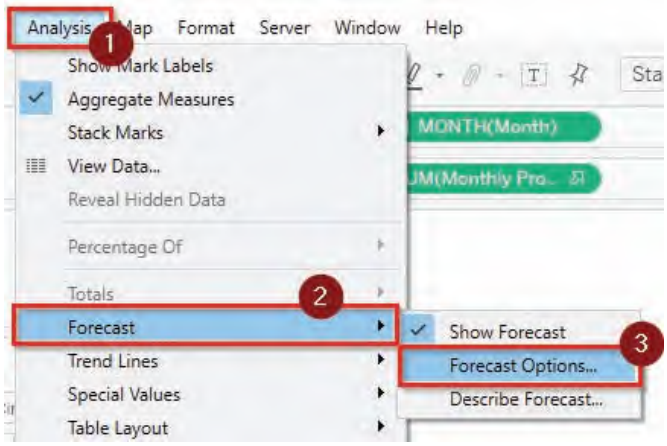
Lab Exhibit 5.4.6

Tableau Software

6. You can also adjust the amount of time you are forecasting. From the Analysis Tab, select **Forecast > Forecast Options** (Lab Exhibit 5.4.7).

Lab Exhibit 5.4.7

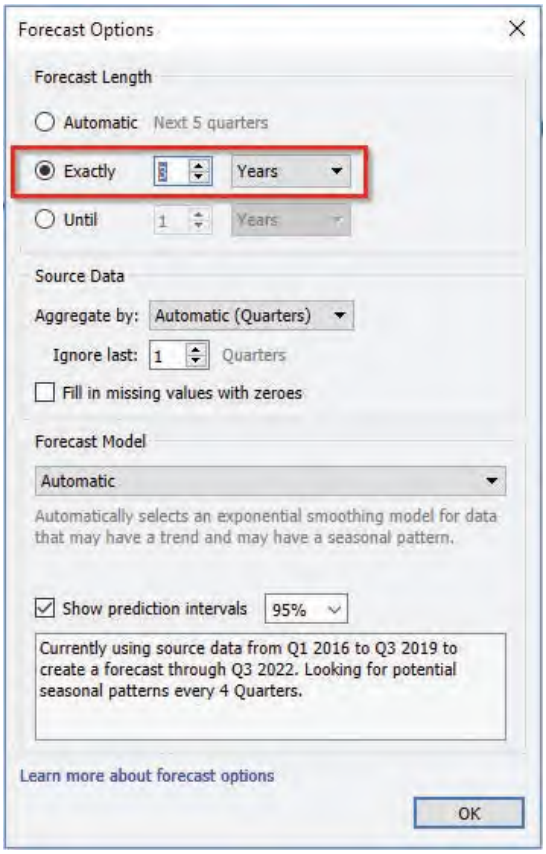
Tableau Software



7. Adjust the Forecast Length to Exactly 3 Years and click OK (Lab Exhibit 5.4.8).

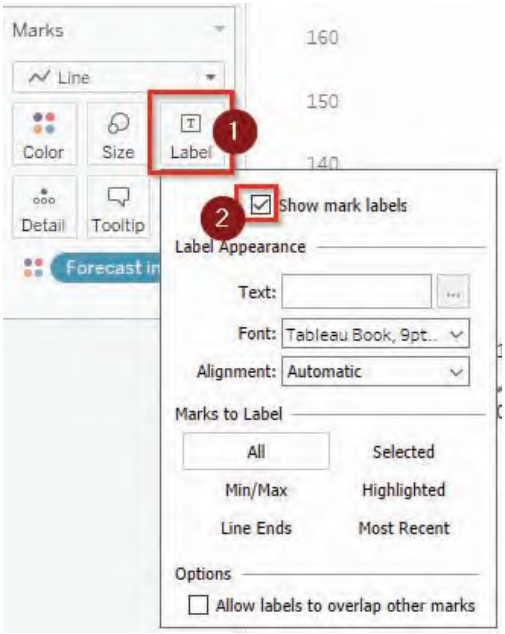
Lab Exhibit 5.4.8

Tableau Software



As you make the change, you can see how the chart adjusts to provide more data. You can also see that the whiskers around the dates extend as you reach further into the future, due to increasing uncertainty about the outcome.

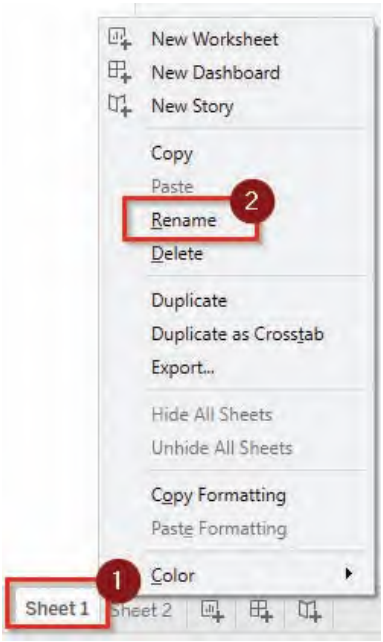
8. Notice that the labels indicating the month and product demand (either actual or estimated) do not appear by default. To add the labels, click Labels in the Marks shelf and add a checkmark next to Show mark labels (Lab Exhibit 5.4.9).



Lab Exhibit 5.4.9

Tableau Software

1. To rename this sheet to "Forecast with Box Plots", right-click on Sheet 1 and select Rename (Lab Exhibit 5.4.10).



Lab Exhibit 5.4.10

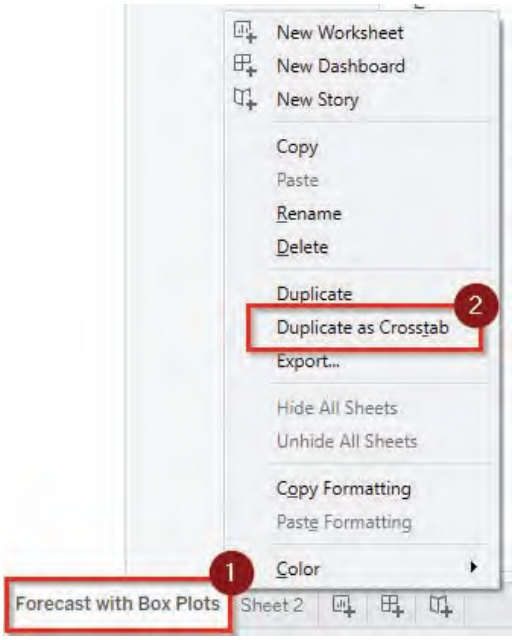
Tableau Software

Analysis Task 3: Create a Table of the Forecast.

1. Tableau does not default to providing the values for the upper and lower prediction levels the way Excel does. Remember, Tableau is used primarily for visualizing data. However, when forecasting, it can be useful to see not only the marks for the predicted values but also the upper and lower prediction levels.
   To see these levels, we need to duplicate our current sheet as a Crosstab. Right-click Forecast with Box Plots and select Duplicate as Crosstab (Lab Exhibit 5.4.11).

Lab Exhibit 5.4.11

Tableau Software



2. You will see a new sheet titled Forecast with Box Plots (2). Rename this sheet to "Forecast crosstab data".
3. Click into your new Forecast with Crosstab data sheet to view the crosstab data. If you scroll to the right, you will see data for upper and lower bounds. The data should appear as shown in Lab Exhibit 5.4.12.

Lab Exhibit 5.4.12

Tableau Software



Forecast Indicator		May 2019	June 2019	July 2019	August 2019	Septemb er 2019	October 2019	Novemb er 2019	Decembe r 2019	January 2020	February 2020	March 2020	April 2020	May 2020	June 2020	July 2020	August 2020	Septemb er 2020	October 2020	Novemb er 2020
Actual	Monthly Pr...	116.74	128.81	130.88	132.19	129.45	129.72	137.30												
Estimate	Lower Predi...								124.46	110.76	120.69	126.80	116.13	118.78	126.00	130.25	133.58	125.92	128.54	135.87
	Monthly Pr...								130.52	116.54	126.63	133.09	122.43	124.69	132.77	137.52	140.10	132.52	135.21	141.98
	Upper Predi...								136.58	122.32	132.57	139.39	128.74	130.60	139.54	144.79	146.63	139.12	141.89	148.09

Report the Results

We have now developed a forecast for future product demand that we can use to plan our operations and sales for the future.

Assessment

1. Take a screenshot of the forecast graph of product demand, paste it into a Word document named "Lab 5.4 Tableau Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the forecast crosstab data (or table) of product demand, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 5.4 Tableau Submission.docx via Connect, if assigned.

Alternate Lab 5.4 Tableau

Forecasting Product Demand Using Time Series Analysis

Required

1. Using Tableau's Forecasting options, forecast product demand (in dollar sales) for Adventure Shoes from 2025 to 2027.
2. Add a duplicate of the sheet to show the cross-tab data.

Open Tableau and connect to Excel file Alt Lab 5.4 Data.xlsx.

Data Dictionary

Month:	Month of the year from 2021 to 2024
Monthly Product Demand:	Product demand (in dollar sales)

Assessment

1. Take a screenshot of the forecast graph of product demand, paste it into a Word document named "Alt Lab 5.4 Tableau Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the forecast crosstab data (or table) of product demand, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 5.4 Tableau Submission.docx via Connect, if assigned.

Lab 5.5 Power BI

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Forecasting Product Demand Using Time Series Analysis

Keywords

Predictive Analysis, Forecasting, Time Series, Product Demand

Decision-Making Context

There are many reasons to forecast a firm's future performance. For example, predicting sales may help a company determine its needed manufacturing capacity. Predicting cash flows may help the company decide if it will need a loan or not. At the individual level, an investor might forecast a company's earnings to evaluate a potential investment.

The goal of this lab is to forecast sales and earnings for Hike 'Em Hiking Boots, a fictitious company.

Required

Using Power BI's forecasting options, forecast product demand for Hike 'Em Hiking Boots from 2025 to 2027 (in dollar sales).

Specify the Question

What will be the monthly demand (in dollar sales) for Hike 'Em Hiking Boots during 2025, 2026, and 2027?

Obtain the Data

1. Open Power BI Desktop. Connect Power BI to the Excel file. From Power BI Desktop, select **Home > Get Data > Excel Workbook**, then browse to the location where you have saved file Lab 5.5 Data.xlsx, and click Open.
2. In the Navigator window, check the box next to Sheet 1 and click Load.

Data Dictionary

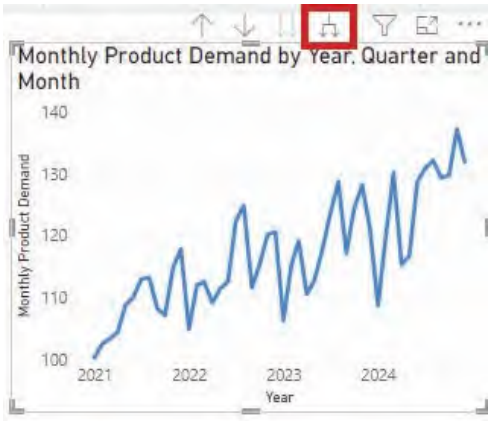
Month:	Month of the year from 2021 to 2024
Monthly Product Demand:	Product demand (dollar sales of boots)

Perform the Analysis

Analysis Task 1: Plot the Time Series Data.

4. To start the forecast analysis, click the Line chart button in the Visualizations pane. Check the boxes next to (1) Date Hierarchy (after expanding MONTH in the fields), to get Date Hierarchy....check year, quarter, month and day, and (2) Monthly Product Demand to add them to your visual as Axis and Values, respectively.

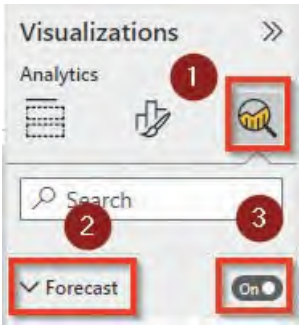
1. To expand the dates to show Months instead of Years, click Expand all down two levels in the hierarchy twice (upside-down fork button in the top-right corner of the visual), as shown in Lab Exhibit 5.5.1.



Lab Exhibit 5.5.1

Microsoft Power BI

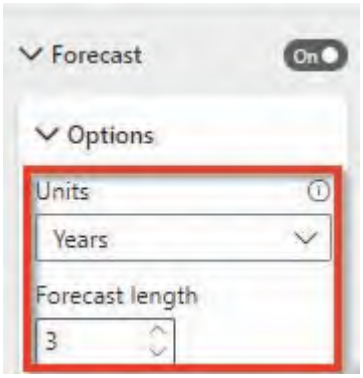
2. To expand this line chart to show a Forecast, click the Analytics magnifying glass button in the Visualizations Pane. Expand the Forecast header and click +Add, as shown in Lab Exhibit 5.5.2.



Lab Exhibit 5.5.2

Microsoft Power BI

3. Adjust the Forecast Length to Exactly 3 Years, as shown in Lab Exhibit 5.5.3. Click Apply.



Lab Exhibit 5.5.3

Microsoft Power BI

Report the Results

We have now developed a forecast for future product demand that we can use to plan our operations and sales for the future.

Assessment

1. Take a screenshot of the forecast graph of product demand, paste it into a Word document named "Lab 5.5 Power BI Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Lab 5.5 Power BI Submission. docx via Connect, if assigned.

Alternate Lab 5.5 Power BI

Forecasting Product Demand Using Time Series Analysis

Required

Using Power BI's Forecasting options, forecast product demand for Adventure Shoes from 2025 to 2027.

Open Tableau and connect to Excel file Alt Lab 5.5 Data.xlsx.

Data Dictionary

Month:	Month of the year from 2021 to 2024
Monthly Product Demand:	Product demand (dollar sales of hiking boots)

Assessment

1. Take a screenshot of the forecast graph of product demand, paste it into a Word document named "Alt Lab 5.5 Power BI Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 5.5 Power BI Submission.docx via Connect, if assigned.

Lab 5.6 Excel

Forecasting Product Demand Using Regression

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Predictive Analysis, Forecasting, Regression, Product Demand

Decision-Making Context

There are many reasons to forecast future product demand. For example, a company may forecast future product demand to make sure it has adequate product available for sale.

Hike 'Em Hiking Boots is a fictitious company that sells hiking boots and works hard to have the right product on hand and available for sale.

In Lab 5.3, we used a time series model (past values of product demand) to predict future demand. In this lab, we use regression analysis to assess if other variables are associated with monthly product demand.

Required

Use regression analysis to perform a cross-sectional analysis to determine what drove Hike 'Em's monthly product demand from 2021 to 2024.

Specify the Question

What were the drivers of the monthly demand (dollar sales of hiking boots) for Hike 'Em Hiking Boots from 2021 to 2024?

Obtain the Data

Open Excel file Lab 5.6 Data.xlsx and browse around to ensure the data are complete.

Data Dictionary

Month Beginning:	The date that the production month begins.
Monthly Product Demand:	The monthly product demand for hiking boots.
GDP:	Gross domestic product, a measure of the overall economy. This number comes from the Federal Reserve (the U.S. central bank).
Weather:	The average temperature (in degrees Fahrenheit) by month of the year.

Analyze the Data

Analysis Task 1: Evaluate Potential Predictors of Monthly Product Demand.

1. We hypothesize that there might be three predictors of monthly product demand for hiking boots. It is a good idea to think through the variables. For example, it would seem that the warmer the weather, the more likely we are to sell hiking boots, because people prefer to go hiking in warmer weather. Although weather is one of the variables we'll evaluate, there are also other possible independent variables. Here are the three independent variables we will evaluate:

   a. **Overall Performance of the Economy.** Would you expect to see greater demand for hiking boots as the economy improves? We expect that as gross domestic product (a commonly used measure of the economy's health) increases, the more hiking boots will be sold. In other words, we expect a positive (+) relationship between Monthly Product Demand and GDP.

b. **Weather (Temperature).** Would you expect to see greater demand for hiking boots as the temperature gets warmer? Are people more likely to hike in the summertime? All other things equal, we expect that the warmer the weather, the more hiking boots will be sold. In other words, we expect a positive (+) relationship between Monthly Product Demand and Weather (Average Temperature Fahrenheit).

c. **Holiday Season.** Would you expect to see greater demand for hiking boots during the holiday season? All other things equal, we expect that during the holiday months (November and December), more hiking boots will be sold. In other words, we expect a positive (+) relationship between Monthly Product Demand and Holiday Season.

We summarize these relationships as follows:

Independent Variable	Expected Relationship with Monthly Product Demand for Hiking Boots
GDP (Gross Domestic Product)	+
Weather (Average Temperature Fahrenheit)	+
Holiday Season	+

Analysis Task 2: Prepare the Data for Analysis.

1. We need to create a holiday variable. To do this, in column E, insert the title "Holiday" and then starting with cell E2, insert a 1 for all months of November and December. Then insert a "0" for all other months for the full data set. The resulting spreadsheet should appear as shown in Lab Exhibit 5.6.1.

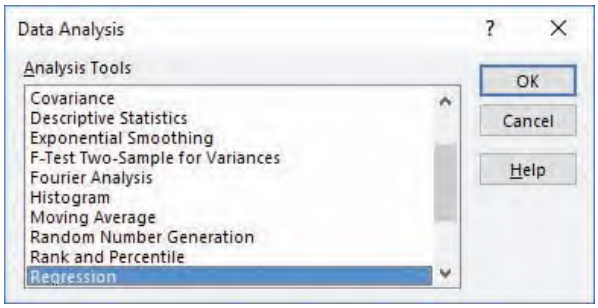
Lab Exhibit 5.6.1

Microsoft Power BI

	A	B	C	D	E
1	Month	Monthly Product Demand	GDP	Weather (Avg Fahrenheit)	Holiday
2	2021-01-01	100.32	100.6424814850660	33	0
3	2021-02-01	102.57	100.8432906500780	35	0
4	2021-03-01	103.32	101.0447723452060	40	0
5	2021-04-01	104.45	101.2469371814550	50	0
6	2021-05-01	108.78	101.4497957889850	60	0
7	2021-06-01	110.10	101.6533584852450	70	0
8	2021-07-01	112.99	101.8576349350620	75	0
9	2021-08-01	113.27	102.0626339917130	73	0
10	2021-09-01	108.22	102.2683635312760	71	0
11	2021-10-01	107.20	102.4748302801500	65	0
12	2021-11-01	114.90	102.6820395415130	55	1
13	2021-12-01	117.88	102.8899949138870	40	1
14	2022-01-01	104.92	103.0986980016700	33	0

Analysis Task 3: Analyze the Data Using Regression Analysis.

1. Using regression analysis, we can assess which of the independent variables impact the monthly product demand, and the size of that impact, by looking at the independent variables that are statistically related to the monthly product demand.
2. To run the regression, we make sure our Analysis ToolPak is loaded by navigating to **Data > Analysis** in the ribbon and seeing if the Data Analysis Add-In has been installed. If it hasn't, go to Appendix E at the end of this book to learn how to install the Data Analysis ToolPak add-in.
3. To perform the regression, select **Data > Analysis > Data Analysis.** A dialog box will open. Select Regression, as shown in Lab Exhibit 5.6.2.

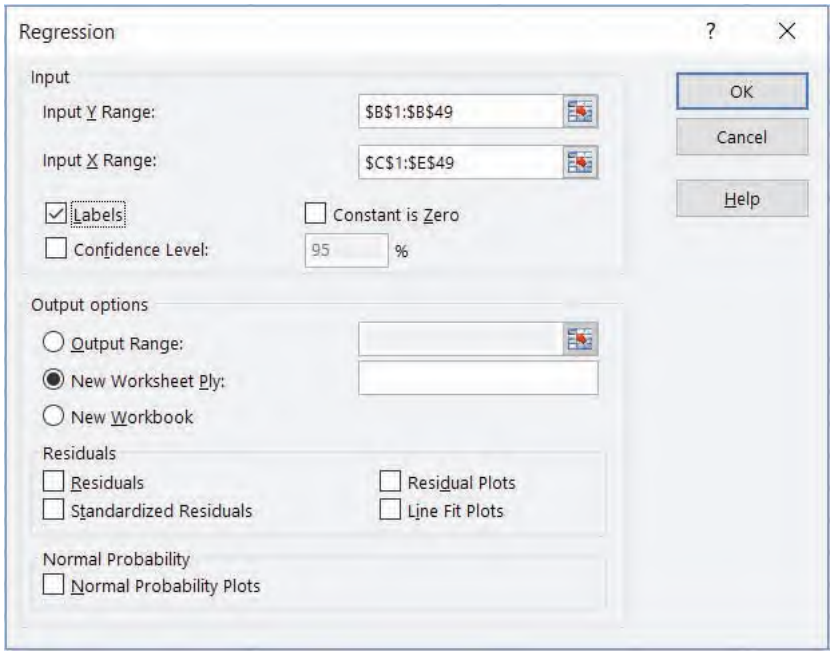


Lab Exhibit 5.6.2

Microsoft Power BI

4. A dialog box will open. We need to select the dependent (outcome) variable (or Y Range). Our dependent (outcome) variable is Monthly Product Demand, so enter "$B$1:$B$49" in the Input Y Range.
5. Next we need to select our independent variables (or X Range). The variables GDP, Weather (Avg Fahrenheit), and Holiday are our independent variables, so enter "$C$1:$E$49" in the Input X Range.
6. Because we highlighted the labels in the data range, click on Labels. Also select New Worksheet Ply to get the results in a new worksheet.

The completed Regression dialog box should look as shown in Lab Exhibit 5.6.3.



Lab Exhibit 5.6.3

Microsoft Power BI

7. Now click OK to get the regression results, which are shown in Lab Exhibit 5.6.4.

Lab Exhibit 5.6.4

Microsoft Power BI

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.891401552				
R Square	0.794596727				
Adjusted R Square	0.780591959				
Standard Error	4.286745807				
Observations	48				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	3127.861777	1042.621	56.73758	3.6376E-15
Residual	44	808.5523431	18.37619		
Total	47	3936.41412			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-123.7576339	22.11431598	-5.59627	1.32E-06	-168.3261093
GDP	2.162221318	0.212246208	10.18733	3.76E-13	1.734467193
Weather (Avg Fahrenheit)	0.202019743	0.043196935	4.676715	2.78E-05	0.114962041
Holiday	7.513785459	1.745077976	4.305702	9.17E-05	3.99681189

There are several things to note about the regression. First, the regression performed better at predicting product demand than a prediction made by random chance. We learn this by looking at the "Significance F" result, which is very small, suggesting there is virtually zero probability that revenues can be explained by no independent variables. This result suggests that we will be able to identify some of the factors that explain product demand.

We can use another statistic, Adjusted R Square, to measure how well these variables predict product demand. *Note:* Statisticians prefer the term "adjusted R-squared," but Excel doesn't use that term. The adjusted R-squared is a value between 0 and 1. An adjusted R-squared value of 0 indicates that the model has no ability to explain the dependent (outcome) variable, and an adjusted R-square value of 1 indicates that the model has perfect ability to explain the dependent (outcome) variable. In this case, the adjusted R-squared value is 0.781, which represents that the model provides reasonably high ability to explain monthly product demand.

The statistics also report that all three independent variables—GDP, Weather, and Holiday Season—help to predict product demand. As the output shows, the t State is greater than two for each of the independent variables, suggesting that these three factors have a positive statistical relationship in explaining and predicting product demand.

Report the Results

We have now developed a forecast for future product demand that we can use to plan our future operations and sales. If we have a prediction of the economy's performance (GDP), the monthly temperature, and whether it is a holiday month or not, we will have good predictors of the monthly product demand for hiking boots. We could also think about whether there are other determinants, or even if a time series analysis (as shown in Lab 5.3) might be better than this forecast of future product demand.

Assessment

1. Take a screenshot of the regression results predicting product demand, paste it into a Word document named "Lab 5.6 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Lab 5.6 Excel Submission.docx via Connect if assigned.

Alternate Lab 5.6 Excel

Forecasting Product Demand Using Regression

Apply the same analysis employed in Lab 5.6 to Alt Lab 5.6 for Adventure Shoes.

Required

Use regression analysis to perform a cross-sectional analysis to determine what drove monthly product demand at Adventure Shoes from 2021 to 2024. Remember to compute the holiday season with "1" in the months of November and December; use 0 for all other months.

Open Excel file Lab Alt Lab 5.6 Data.xlsx.

Data Dictionary

Month Beginning:	The date that the production month begins.
Monthly Product Demand:	The monthly product demand for hiking boots.
GDP:	Gross domestic product, a measure of the overall economy. This number comes from the Federal Reserve (the U.S. central bank).
Weather:	The average temperature (in degrees Fahrenheit) by month of the year.

Assessment

1. Take a screenshot of the regression results predicting product demand, paste it into a Word document named "Alt Lab 5.6 Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 5.6 Submission.docx via Connect, if assigned.

Lab 5.7 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Assessing the Returns to Education: To MBA or Not to MBA?

Keywords

NPV(), IRR(), Discounting Cash Flows

Decision-Making Context

You are nearing the end of your senior year in college, and you are trying to decide if you should get an MBA or take a job and start your career. While there are many possible decision criteria, we'll explore the financial component. That is, which course of action will give you the higher long-term payoff or the higher net present value?

Required

1. Determine the net present value of earnings from a job after receiving an undergraduate degree and compare it with the net present value of earnings from a job after getting an MBA degree given various assumptions.
2. Create a visualization of the difference in cash flows from earnings after getting or not getting an MBA degree.

Specify the Question

What is the net present value of your earnings from taking a job after your undergraduate degree as compared to the net present value of your earnings after getting an MBA degree?

Obtain the Data

We know that the amounts, timing, and uncertainty of future cash flows have implications on the net present value. We will develop our data from scratch.

Open Lab 5.7 Data.xlsx and browse its contents. It will serve as a template for our analysis.

Analyze the Data

Analysis Task 1: Determine the Net Present Value for the First Alternative: Getting an MBA.

1. In cells B4:B6, insert the number "– 20000", representing the net cost of getting an MBA.
2. In cell B7, insert the number "75000", representing the first year's salary with an MBA-type job.
3. In cell B8, insert the formula "=B7*1.03", representing a 3% raise.
4. Copy the contents of cell B8 to cells B9:B39, representing a 3% raise each year.
5. Select cells B4:B39, right-click, and format the cells to be currency with two decimal places.
6. We now need to discount these cash flows. By using the =NPV() function, we can compute the net present value of these cash flows. To do so, we need to specify the following:
   =NPV(rate, string of cash flows) where
   rate = the annual interest or discount rate, which is 9% (or .09) in this example; and
   string of cash flows = the cash flows that we expect to pay or be paid each year.

In cell B2, insert the formula "=NPV(0.09,B4:B39)", as shown in Lab Exhibit 5.7.1. Cells B4:B39 represent the string of cash flows from age 25 to 60. This formula will calculate the present value of this string of cash flows.

Lab Exhibit 5.7.1

	A	B
1		Getting an MBA
2	NPV:	=NPV(0.09,B4:B39)
3	Age	
4	25	-$20,000.00
5	26	-$20,000.00
6	27	-$20,000.00
7	28	$75,000.00
8	29	$77,250.00
9	30	$79,567.50
10	31	$81,954.53
11	32	$84,413.16
12	33	$86,945.56
13	34	$89,553.92
14	35	$92,240.54
15	36	$95,007.76
16	37	$97,857.99
17	38	$100,793.73
18	39	$103,817.54
19	40	$106,932.07
20	41	$110,140.03
21	42	$113,444.23
22	43	$116,847.56
23	44	$120,352.98
24	45	$123,963.57
25	46	$127,682.48
26	47	$131,512.95
27	48	$135,458.34
28	49	$139,522.09
29	50	$143,707.76
30	51	$148,018.99
31	52	$152,459.56
32	53	$157,033.34
33	54	$161,744.35
34	55	$166,596.68
35	56	$171,594.58
36	57	$176,742.41
37	58	$182,044.69
38	59	$187,506.03
39	60	$193,131.21

The result is as shown in Lab Exhibit 5.7.2.

Lab Exhibit 5.7.2

	A	B
1		Getting an MBA
2	NPV:	$765,604.06
3	Age	

The NPV analysis suggests that with the string of cash flows after receiving an MBA discounted at 9%, the net present value is $765,604.06.

7. We'd like to check that analysis, so we'll do it a second way. We know that to discount, we use the following equation:

$$\text{Present Value of Year } t\text{'s cash flow: } \frac{CF_t}{(1+r)^t}$$

8. First, we calculate *t* for each year. Starting in cell D4, insert the number "1". Then add 1 each year until we put 36 in year cell D39. Column D should appear as shown in Lab Exhibit 5.7.3.

Lab Exhibit 5.7.3

	D
1	
2	
3	
4	1
5	2
6	3
7	4
8	5
9	6
10	7
11	8
12	9
13	10
14	11
15	12
16	13
17	14
18	15
19	16
20	17
21	18
22	19
23	20
24	21
25	22
26	23
27	24
28	25
29	26
30	27
31	28
32	29
33	30
34	31
35	32
36	33
37	34
38	35
39	36

9. Insert the equation "=B4/1.09^D4" in cell C4. Then copy that result from cell C4 to cells C5:C39. The results are shown in Lab Exhibit 5.7.4.
10. Select cells C5:C39, right-click, and format the cells to be currency with two decimal places.

Lab Exhibit 5.7.4

	A	B	C	D
1		Getting an MBA		
2	NPV:	$765,604.06		
3	Age		$765,604.06	
4	25	-$20,000.00	-$18,348.62	1
5	26	-$20,000.00	-$16,833.60	2
6	27	-$20,000.00	-$15,443.67	3
7	28	$75,000.00	$53,131.89	4
8	29	$77,250.00	$50,207.20	5
9	30	$79,567.50	$47,443.50	6
10	31	$81,954.53	$44,831.93	7
11	32	$84,413.16	$42,364.12	8
12	33	$86,945.56	$40,032.15	9
13	34	$89,553.92	$37,828.54	10
14	35	$92,240.54	$35,746.24	11
15	36	$95,007.76	$33,778.56	12
16	37	$97,857.99	$31,919.19	13
17	38	$100,793.73	$30,162.17	14
18	39	$103,817.54	$28,501.86	15
19	40	$106,932.07	$26,932.95	16
20	41	$110,140.03	$25,450.41	17
21	42	$113,444.23	$24,049.47	18
22	43	$116,847.56	$22,725.64	19
23	44	$120,352.98	$21,474.69	20
24	45	$123,963.57	$20,292.60	21
25	46	$127,682.48	$19,175.57	22
26	47	$131,512.95	$18,120.04	23
27	48	$135,458.34	$17,122.60	24
28	49	$139,522.09	$16,180.08	25
29	50	$143,707.76	$15,289.43	26
30	51	$148,018.99	$14,447.81	27
31	52	$152,459.56	$13,652.52	28
32	53	$157,033.34	$12,901.00	29
33	54	$161,744.35	$12,190.86	30
34	55	$166,596.68	$11,519.80	31
35	56	$171,594.58	$10,885.68	32
36	57	$176,742.41	$10,286.47	33
37	58	$182,044.69	$9,720.24	34
38	59	$187,506.03	$9,185.18	35
39	60	$193,131.21	$8,679.58	36

11. In cell C3, will sum the individual cash flows. To do so, insert the formula "=SUM(C4:C39)". We find the result is the same $765,604.06 that we calculated using the =NPV formula.

The analysis suggests that given our assumptions, the net present value of cash flows accruing from the decision date is $765,604.

Analysis Task 2: Determine the Net Present Value for the Second Alternative: Not Getting an MBA.

1. We'll now compute the cash flows of not getting an MBA but rather taking a $45,000 job immediately after receiving an undergraduate degree.
   In cell G4, insert the number "45000", representing the first-year salary.
2. Because we assume the salary will increase at 2% per year, in cell G5 insert the formula "=G4*1.02".
3. Copy the result from cell G5 to cells G6:G39 to arrive at the results shown in Lab Exhibit 5.7.5.
4. Select cells G5:G39, right-click, and format the cells to be currency with two decimal places.

Lab Exhibit 5.7.5

F	G
Not getting an MBA	
NPV:	$583,923.03
Age	
25	$45,000.00
26	$45,900.00
27	$46,818.00
28	$47,754.36
29	$48,709.45
30	$49,683.64
31	$50,677.31
32	$51,690.86
33	$52,724.67
34	$53,779.17
35	$54,854.75
36	$55,951.84
37	$57,070.88
38	$58,212.30
39	$59,376.54
40	$60,564.08
41	$61,775.36
42	$63,010.86
43	$64,271.08
44	$65,556.50
45	$66,867.63
46	$68,204.99
47	$69,569.09
48	$70,960.47
49	$72,379.68
50	$73,827.27
51	$75,303.82
52	$76,809.89
53	$78,346.09
54	$79,913.01
55	$81,511.27
56	$83,141.50
57	$84,804.33
58	$86,500.41
59	$88,230.42
60	$89,995.03

5. In cell G2, we calculate the net present value of not getting an MBA. To do so, insert the NPV function, "=NPV(0.09,G4:G39)". We calculate a NPV of $583,923, as shown in Lab Exhibit 5.7.6.

Lab Exhibit 5.7.6

	A	B	C	D	E	F	G
1		Getting an MBA				Not getting an MBA	
2	NPV:	$765,604.06				NPV:	$583,923.03
3	Age		$765,604.06			Age	
4	25	-$20,000.00	-$18,348.62	1		25	$45,000.00
5	26	-$20,000.00	-$16,833.60	2		26	$45,900.00
6	27	-$20,000.00	-$15,443.67	3		27	$46,818.00
7	28	$75,000.00	$53,131.89	4		28	$47,754.36
8	29	$77,250.00	$50,207.20	5		29	$48,709.45
9	30	$79,567.50	$47,443.50	6		30	$49,683.64
10	31	$81,954.53	$44,831.93	7		31	$50,677.31
11	32	$84,413.16	$42,364.12	8		32	$51,690.86
12	33	$86,945.56	$40,032.15	9		33	$52,724.67
13	34	$89,553.92	$37,828.54	10		34	$53,779.17
14	35	$92,240.54	$35,746.24	11		35	$54,854.75
15	36	$95,007.76	$33,778.56	12		36	$55,951.84
16	37	$97,857.99	$31,919.19	13		37	$57,070.88
17	38	$100,793.73	$30,162.17	14		38	$58,212.30
18	39	$103,817.54	$28,501.86	15		39	$59,376.54
19	40	$106,932.07	$26,932.95	16		40	$60,564.08
20	41	$110,140.03	$25,450.41	17		41	$61,775.36
21	42	$113,444.23	$24,049.47	18		42	$63,010.86
22	43	$116,847.56	$22,725.64	19		43	$64,271.08
23	44	$120,352.98	$21,474.69	20		44	$65,556.50
24	45	$123,963.57	$20,292.60	21		45	$66,867.63
25	46	$127,682.48	$19,175.57	22		46	$68,204.99
26	47	$131,512.95	$18,120.04	23		47	$69,569.09
27	48	$135,458.34	$17,122.60	24		48	$70,960.47
28	49	$139,522.09	$16,180.08	25		49	$72,379.68
29	50	$143,707.76	$15,289.43	26		50	$73,827.27
30	51	$148,018.99	$14,447.81	27		51	$75,303.82
31	52	$152,459.56	$13,652.52	28		52	$76,809.89
32	53	$157,033.34	$12,901.00	29		53	$78,346.09
33	54	$161,744.35	$12,190.86	30		54	$79,913.01
34	55	$166,596.68	$11,519.80	31		55	$81,511.27
35	56	$171,594.58	$10,885.68	32		56	$83,141.50
36	57	$176,742.41	$10,286.47	33		57	$84,804.33
37	58	$182,044.69	$9,720.24	34		58	$86,500.41
38	59	$187,506.03	$9,185.18	35		59	$88,230.42
39	60	$193,131.21	$8,679.58	36		60	$89,995.03

Report the Results

Many factors help people decide whether to get an MBA (or other graduate degree) or not. Determining the financial implications is one important factor. Knowing the net present value of each alternative helps us determine if getting an MBA is worth the investment in time and money.

How could you visualize the results to understand them better and compare the annual or cumulative dollar figures over your lifetime?



DATA VISUALIZATION

1. If we take one path, getting an MBA, we forego the opportunity of getting a job immediately. To help visualize the results, in cell K4, insert the equation "=-G4+B4". This equation helps us determine the difference between the cash flows associated with an MBA and the cash flows associated with not getting an MBA.

2. Copy the result from cell K4 to cells K5:39.
3. Highlight cells K4:K39, and insert a chart by selecting **Insert > Charts > 2D Clustered Column** Chart to view the visualization.
4. Label the chart "Incremental Value of Getting an MBA".

Assessment

1. Take a screenshot of the completed spreadsheet, paste it into a Word document named "Lab 5.7 Excel Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the visualization you created to see the difference in cash flows from earnings between getting an MBA and not getting an MBA, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 5.7 Excel Submission.docx via Connect, if assigned.

Alternate Lab 5.7 Excel

Assessing the Returns to Education: To MBA or Not to MBA?

Apply the same steps in Lab 5.7 to the Alt Lab 5.7 Data.xlsx template.

Required

1. Determine the net present value of taking a job after receiving an undergraduate degree and compare/contrast it with the net present value of getting an MBA degree given various assumptions.
2. Create a visualization of the difference in cash flows between getting and not getting an MBA degree.

   Note the differences between Alt Lab 5.7 and Lab 5.7. All of the assumptions from Lab 5.7 hold, except the following:
   The starting salary if you get an MBA is $90,000 and the discount rate is 6%.
   The starting salary if you do not get an MBA is $55,000 and the discount rate is 6%.

Assessment

1. Take a screenshot of the completed spreadsheet, paste it into a Word document named "Alt Lab 5.7 Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the visualization you created to see the difference in cash flows between getting an MBA and not getting an MBA, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 5.7 Submission.docx via Connect, if assigned.

Lab 5.8 Excel

Applying Scenario Analysis: Possible Trade War

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

What-if Scenarios, Trade War Impact

Decision-Making Context

What-if scenario analysis considers the potential outcomes of possible future events. It does not try to predict one possible outcome, but rather presents a range of alternative outcomes.

As an example, assume there is a possibility of a trade war with a major trading partner that may have no effect (no trade war), a 10% decrease in income, or a 25% decrease in income. At the same time (and unrelated to the trade war), a company might have a potential increase in domestic demand that will be neutral, cause income to increase 5%, or cause income to decrease by 5%. The CFO has asked you to evaluate the expected potential impact on income of all these potential scenarios at once.

Required

Calculate the expected dollar impact of the various potential scenarios based on the probability of each scenario occurring.

Specify the Question

What is the expected income impact of a possible trade war?

Obtain the Data

Iphone Coverage is a fictitious small company that manufactures and sells covers for iPhones. The company is evaluating the impact of a potential trade war on its profits.

Open Excel file Lab 5.8 Data.xlsx, which provides a template for our analysis. The data file should appear as shown in Lab Exhibit 5.8.1.

Lab Exhibit 5.8.1

Income Level for Each of These Scenarios				
		No Trade War	Moderate Trade War	Severe Trade War
Income Scenarios	Income	0%	-10%	-25%
Domestic Demand Increases Income by 5%	2,100,000			
Domestic Demand Doesn't Change	2,000,000			
Domestic Demand Decreases Income by 5%	1,900,000			

Analyze the Data

Analysis Task 1: Set Up the Nine Different Possible Scenarios Based on Trade War Status and Domestic Demand.

All in all, there are nine possible scenarios measuring the impact of the possible trade war. If there is no trade war, there will be no impact on the company's net income. There will be a 10% decrease in income if there is a moderate trade war and a 25% decrease in income if there is a severe trade war.

1. Using cell B5 as the base case, calculate the change in income from the base case ($B$5) for each of the nine scenarios using the formulas in cells C4:E9 and in cells C13:E15, as shown in Lab Exhibit 5.8.2.

Lab Exhibit 5.8.2

	A	B	C	D	E
1	**Income Level for Each of These Scenarios**				
2			No Trade War	Moderate Trade War	Severe Trade War
3	Income Scenarios	Income	0	-0.1	-0.25
4	Domestic Demand Increases Income by 5%	=(1+0.05)*B5	=$B$5*(1+0.05)*(1-$C$3)	=$B$5*(1+0.05)*(1+$D$3)	=$B$5*(1+0.05)*(1+$E$3)
5	Domestic Demand Doesn't Change	2000000	=$B$5*1*(1-$C$3)	=$B$5*1*(1+$D$3)	=$B$5*1*(1+$E$3)
6	Domestic Demand Decreases Income by 5%	=(1-0.05)*B5	=$B$5*(1-0.05)*(1+$C$3)	=$B$5*(1-0.05)*(1+$D$3)	=$B$5*(1-0.05)*(1+$E$3)
7					
8					
9	**The change in income based on scenario above:**				
10					
11			No Trade War	Moderate Trade War	Severe Trade War
12	Income Scenarios	Income	0	-0.1	-0.25
13	Domestic Demand Increases Income by 5%	=(1+0.05)*B14	=C4-$B$5	=D4-$B$5	=E4-$B$5
14	Domestic Demand Doesn't Change	2000000	=C5-$B$5	=D5-$B$5	=E5-$B$5
15	Domestic Demand Decreases Income by 5%	=(1-0.05)*B14	=C6-$B$5	=D6-$B$5	=E6-$B$5

The result is shown in Lab Exhibit 5.8.3.

Lab Exhibit 5.8.3

	A	B	C	D	E
1	**Income Level for Each of These Scenarios**				
2			No Trade War	Moderate Trade War	Severe Trade War
3	Income Scenarios	Income	0%	-10%	-25%
4	Domestic Demand Increases Income by 5%	2,100,000	2,100,000	1,890,000	1,575,000
5	Domestic Demand Doesn't Change	2,000,000	2,000,000	1,800,000	1,500,000
6	Domestic Demand Decreases Income by 5%	1,900,000	1,900,000	1,710,000	1,425,000
7					
8					
9	**The change in income based on scenario above:**				
10					
11			No Trade War	Moderate Trade War	Severe Trade War
12	Income Scenarios	Income	0%	-10%	-25%
13	Domestic Demand Increases Income by 5%	2,100,000	100,000	(110,000)	(425,000)
14	Domestic Demand Doesn't Change	2,000,000	-	(200,000)	(500,000)
15	Domestic Demand Decreases Income by 5%	1,900,000	(100,000)	(290,000)	(575,000)

2. Next, we need to estimate the probability of each scenario occurring. After talking to management, we have determined the expected probability of each scenario, which are given in the table below. Input the appropriate values in cells C20:E22, as shown in Lab Exhibit 5.8.4. The total of the nine cells will equal 100%, so make sure you have entered each cell's probability correctly.

Lab Exhibit 5.8.4

	Taxable Income	No Trade War	Moderate Trade War	Severe Trade War
Domestic Demand Increases Income by 5%	2,100,000	0.05	0.20	0.25
Domestic Demand Doesn't Change	2,000,000	0.05	0.15	0.10
Domestic Demand Decreases Income by 10%	1,900,000	0.10	0.05	0.05

Analysis Task 2: Determine the Expected Net Income Impact of Each of the Nine Scenarios and Overall Impact.

1. We are now ready to calculate the expected impact of each individual scenario. We do so by multiplying the expected dollar change in income by the probability of that scenario occurring. For example, in cell C28, multiply cell C13 by cell C20.
2. We then extend that calculation to all cells. Multiply the expected dollar change in income by its respective probability in the grid below and then sum the expected impact by row and by column and overall to come up with the result shown in Lab Exhibit 5.8.5.

Lab Exhibit 5.8.5

Income Level for Each of These Scenarios					
		No Trade War	Moderate Trade War	Severe Trade War	
Income Scenarios	Income	0%	-10%	-25%	
Domestic Demand Increases	2,100,000	2,100,000	1,890,000	1,575,000	
Domestic Demand Doesn't C	2,000,000	2,000,000	1,800,000	1,500,000	
Domestic Demand Decreases	1,900,000	1,900,000	1,710,000	1,425,000	
The change in income based on scenario above:					
		No Trade War	Moderate Trade War	Severe Trade War	
Income Scenarios	Income	0%	-10%	-25%	
Domestic Demand Increases	2,100,000	100,000	(110,000)	(425,000)	
Domestic Demand Doesn't C	2,000,000	-	(200,000)	(500,000)	
Domestic Demand Decreases	1,900,000	(100,000)	(290,000)	(575,000)	
Probability of the scenario occurring:					
		No Trade War	Moderate Trade War	Severe Trade War	
Income Scenarios	Income	0%	-10%	-25%	Sum
Domestic Demand Increases	2,100,000	0.05	0.20	0.25	0.50
Domestic Demand Doesn't C	2,000,000	0.05	0.15	0.10	0.30
Domestic Demand Decreases	1,900,000	0.10	0.05	0.05	0.20
Sum		0.20	0.40	0.40	1.00
Expected impact of changes in income due to trade war					
		No Trade War	Moderate Trade War	Severe Trade War	
Income Scenarios	Income	0%	-10%	-25%	Sum of Expected Savings
Domestic Demand Increases	2,100,000	5,000	(22,000)	(106,250)	(123,250)
Domestic Demand Doesn't C	2,000,000	-	(30,000)	(50,000)	(80,000)
Domestic Demand Decreases	1,900,000	(10,000)	(14,500)	(28,750)	(53,250)
Sum of Expected Change in		(5,000)	(66,500)	(185,000)	(256,500)

3. According to our scenario analysis, the expected impact of the trade war on income is a decrease in income of $256,500. We get this number by summing the nine cells of expected impact, or the row or column totals. Management can use this information to evaluate the expected impact of a trade war on profitability.

Report the Results

Using what-if scenario analysis, Iphone Coverage managers are able to estimate an expected decrease in income of $256,500. This information will help them determine how many products to produce, how to manage the company's workforce, and how to manage the company's cash balances.

Assessment

1. Take a screenshot of the completed spreadsheet, paste it into a Word document named "Lab 5.8 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Lab 5.8 Submission.docx via Connect, if assigned.

Alternate Lab 5.8

Applying Scenario Analysis: Possible Trade War

Peters Engineering (a fictitious company) has income of $3,000,000 with a potential increase or decrease of 10% due to domestic demand. Evaluate the expected change in income based on a possibility of no trade war, moderate trade war (decrease in income of 5%), and severe trade war (decrease in income of 30%).

The following table summarizes the probability of each trade-war scenario and change in domestic demand occurring:

	Taxable Income	No Trade War	Moderate Trade War	Severe Trade War
Domestic Demand Increases Income by 10%	3,300,000	0.10	0.05	0.10
Domestic Demand Doesn't Change	3,000,000	0.10	0.20	0.10
Domestic Demand Decreases Income by 10%	2,700,000	0.15	0.05	0.15

Apply the same steps in Lab 5.8 to the Alt Lab 5.8.xlsx dataset.

Assessment

1. Take a screenshot of the completed spreadsheet, paste it into a Word document named "Alt Lab 5.8 Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 5.8 Submission.docx via Connect, if assigned.

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Chapter 6

Report the Results:
Using Data Visualization



DATA VISUALIZATION

A Look Back

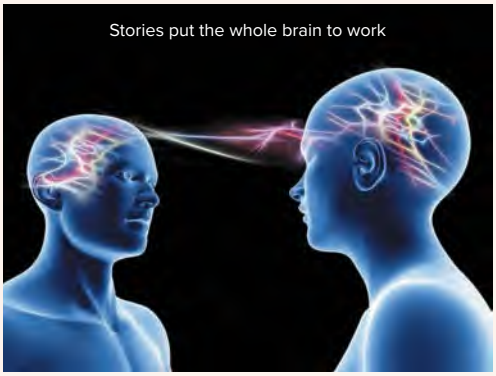
- Chapters 4 and 5 explained the fourth step of the SOAR analytics model, ANALYZE THE DATA.
- Chapter 4 introduced exploratory business analytics, or the initial descriptive and diagnostic analytical investigations to summarize and explain performance.
- Chapter 5 introduced confirmatory business analytics, or predictive and prescriptive analytical investigations. Predictive analytics forecasts or predicts the future (for example, forecasting or predicting product demand). Prescriptive analytics optimizes the company's response to the projected future.

A Look at This Chapter

- The final step of the SOAR analytics model is REPORT THE RESULTS.
- Results can be shared using a variety of different types of reports and visualizations that answer the original business question(s).
- The appropriate visualization depends on the type of analytics, either exploratory or confirmatory.
- To provide better insights, data visualizations can be combined on a dashboard.
- Under certain conditions, written reports can be useful to decision-makers.

A Look Ahead

- Chapters 7 through 10 describe how companies use business analytics to gain actionable business intelligence (BI) across the four main business functions: marketing, accounting, finance, and operations. These four chapters focus on the first four types of analytics: descriptive, diagnostic, predictive, and prescriptive.
- Marketing helps companies meet customer needs profitably by selecting an appropriate marketing mix, which includes product, price, place, and promotion. Chapter 7 investigates how business analysts in the marketing function apply the SOAR model.



Telling a Good Story Visually

Source: https://www.youtube.com/watch?v=iWouh2jGxDo, Accessed February 3, 2022.

The Importance of Good Storytelling

"Stories are up to 22 times more memorable than facts or figures alone."

—STANFORD GRADUATE SCHOOL OF BUSINESS

Think back to all of the conversations or class lectures in which you have participated over your life. You undoubtedly found yourself more engaged by good storytellers, and you probably found it easier to understand information if pictures accompanied it.

For these reasons, companies often use stories in advertisements with the goal of capturing consumers' attention. A good example is **Anheuser-Busch**'s Super Bowl commercials. Often, these commercials use beautiful images and very few words to tell a heartwarming story about the company's famous Budweiser Clydesdale horses. One such commercial follows a Dalmatian training a Clydesdale to become a Budweiser horse. Such visual stories make an impact because:

- 90 percent of the information transmitted to the brain is visual.
- The brain processes visuals 60,000 times faster than it processes text.
- People retain 95 percent of visual messages compared to 10 percent of text messages.
- Stories activate multiple parts of the brain, increasing attention, recall accuracy, and connection with the story and storyteller.

Given how our brain works, it is not surprising that 91 percent of people prefer visual content over written content. For example, on **Facebook**, photos have an interaction rate of 87 percent, compared to 4 percent or less for other types of posts, such as links or text. In addition, 65 percent of Facebook users use it to view photos, and 35.6 percent of all posts are photos. In fact, 147,000 photos, on average, are uploaded to Facebook every minute, with over 350 billion pictures uploaded so far. This emphasis on photos is not limited to Facebook. Viewing photos is also the number-one activity of users on **Instagram**, **Pinterest**, and **Snapchat**.

Visuals also play an important role in the business world. Companies are drowning in a massive and growing tidal wave of data as the amount of data created, replicated, and consumed each year continues to increase. To make sense of this vast amount of data, companies are using business analytics and data visualizations to answer business questions. In this chapter, you will learn how to use data visualizations, which are essentially pictures of data, to report the results of business analytics clearly and concisely.[1]

LEARNING OBJECTIVES

After reading this chapter you should be able to:

- **LO 6.1** Describe the basics of data visualization.
- **LO 6.2** Distinguish among chart types.
- **LO 6.3** Describe appropriate visualizations for exploratory business analytics.
- **LO 6.4** Describe appropriate visualizations for confirmatory business analytics.
- **LO 6.5** Explain dashboards and their benefits.
- **LO 6.6** Explain how to communicate your analysis results in a written summary or report.

LABS FOR CHAPTER 6

LAB 6.1 **Excel:** Descriptive Analytics: Visualizing Pivot Table Data Using Conditional Formatting and Sparklines
LAB 6.2 **Tableau:** Descriptive Analytics: Analyzing Sales Revenue by Product with a Tree Map
LAB 6.2 **Power BI:** Descriptive Analytics: Analyzing Sales Revenue by Product with a Tree Map
LAB 6.3 **Tableau:** Descriptive Analytics: Analyzing Sales Revenue by Customer with a Bar Chart and Filters
LAB 6.3 **Power BI:** Descriptive Analytics: Analyzing Sales Revenue by Customer with a Bar Chart and Filters
LAB 6.4 **Excel:** Descriptive Analytics: Creating a Dashboard Using Pivot Tables and Slicers
LAB 6.5 **Tableau/Power BI:** Descriptive Analytics: Creating a Dashboard

LO 6.1

Describe the basics of data visualization.

6.1 THE BASICS OF DATA VISUALIZATION

The fourth and final step of the SOAR model is REPORT THE RESULTS. This step focuses on synthesizing the findings of business analytics for decision-makers. Business analytics is effective only if the findings are clearly communicated in an appropriate format, summarization level, and language for the target audience. Often, reporting includes data visualizations that make the data understandable. **Data visualization** presents the results of the analysis in an accessible way using charts or graphs.

Why Use Visualizations?

"Visualization gives you answers to questions you didn't know you had."

—Ben Schneiderman, Professor, University of Maryland

Suppose you are interning at a large company and your manager asks you to provide information regarding the states in which the company's customers live. Would you simply point your boss to the Customers table in the sales database? Or would you go a step further and isolate the attributes to the Customer Name and the State? Perhaps you might even run a quick query or create a pivot table to count the number of company's customers in each state.

The best way to answer your manager's question is to provide a short written summary of the answer to the business question, along with a chart that visualizes the results. When the results of business analytics are visualized appropriately, everyone finds the results faster and easier to interpret.

Think back to some of the first data visualizations and categorizations you encountered, such as the nutrition/food pyramid and the periodic table of the elements. Also take a minute to think about how frequently *infographics* (information graphics) are used on social media to present complicated information. Infographics are just another name for data visualizations that present data in easy-to-read charts and graphs with minimal text. All infographics have the same goal: to break down difficult concepts into categories and visual components that are easy to read and understand.

Data visualizations are also important because data sets that appear similar may actually be quite different. For example, two data sets may have the same statistical properties, such as the same mean and the same standard deviation, but when you plot the data you may find that the two data sets look completely different. Without creating a data visualization, such as a scatterplot, you might never know that the data sets are composed of very different data, each with its own story and interpretation. In other words, combining statistical analysis with data visualization provides a more complete picture of the data.

[1]Harry Eisenberg, "Humans Process Visual Data Better," September 15, 2014, https://www.t-sciences.com/news/humans-process-visual-data-better\#:~:text=Visualization%20works%20from%20a%20human,to%20the%20brain%20is%20visual. Accessed 12/21/2021; "Six Reasons to Embrace Visual Commerce in 2018," https://www.pixlee.com/blog/six-reasons-to-embrace-visual-commerce-in-2018/. Accessed 12/21/2021; FinancesOnline, "53 Important Statistics About How Much Data Is [*sic*] Created Every Day," https://financesonline.com/how-much-data-is-created-every-day/. Accessed 12/21/2021; Jenn Chen, February 17, 2021; "20 Facebook Stats to Guide Your 2021 Facebook Strategy," sproutsocial, https://sproutsocial.com/insights/facebook-stats-for-marketers/. Accessed 12/21/2021; Mansoor Iqbal, November 12, 2021; "Facebook Revenue and Usage Statistics (2021)," https://www.businessofapps.com/data/facebook-statistics/. Accessed 12/21/2021; Pam Rutledge, "How Storytelling Affects the Brain," available online at pamelarutledge.com. Accessed 2/3/2022; VMI Studio, untitled infographic available online at https://vmistudio.com/wp-content/uploads/fly-images/18271/why-visual-marketing-is-king-1-e1530545463708-850x632.gif. Accessed 2/3/2022.

SOARing to Success

Data Visualization and Anscombe's Quartet

The idea that data sets may have the same statistical properties but completely different data plots has a name: Anscombe's Quartet. In 1973, Francis J. Anscombe, an English statistician and the founding chairman of the statistics department at Yale University, published a paper titled "Graphs in Statistical Analysis." He wrote the paper because statistics textbooks at the time ignored graphs, focusing solely on mathematical calculations. The paper showed that four different data sets with identical statistical properties (means, variances, correlations, and regressions) had completely different plots. Thus, Anscombe argued, data visualizations can help users better understand the structure of a data set, including outliers. Anscombe therefore recommended using both statistics and graphs to provide readers with a more complete understanding of the data.[2]

Excel, Power BI, and Tableau all have excellent data visualization capabilities. If you conduct most of your analysis in Excel and you want to add a chart to supplement your numerical analysis, continuing to use Excel to visualize your data is a great option. You could also import the Excel data into Tableau and Power BI to create a single chart or a dashboard filled with many useful visualizations. Tableau and Power BI offer additional features, such as handling and visualizing large data sets from a variety of sources.

Let's now look at one way to visualize a simple data set. We want to answer these questions:

- Which company has the highest sales?
- Which company has the lowest sales?

Exhibit 6.1 presents a simple data set of the total sales for five companies.

Exhibit 6.1
Data Series of Total Sales by Company

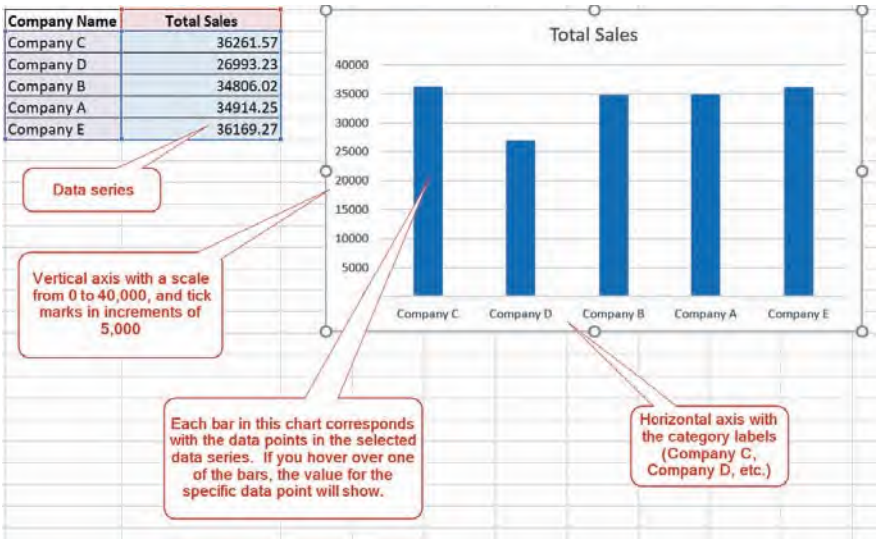
Company Name	Total Sales
Company C	36261.57
Company D	26993.23
Company B	34806.02
Company A	34914.25
Company E	36169.27

Note that this data set doesn't appear in any logical order. The company names are not alphabetized, and the total sales are not sorted in descending or ascending order. Even though the table includes only five companies, it takes a bit of time to read through the numerical values to determine which company has the highest total sales.

[2]F. J. Anscombe, (1973). "Graphs in Statistical Analysis" *The American Statistician*, *27*(1), https://www.sjsu.edu/faculty/gerstman/StatPrimer/anscombe1973.pdf. Accessed 12/23/2021.

It is faster and easier to identify the companies that have the highest and lowest sales if we present the data in a bar chart. Exhibit 6.2 shows the same information for company total sales in chart form, pointing to some key features of the visual. (*Note:* The word *chart* is a general term. Tables, graphs, and diagrams are different types of charts.)

Exhibit 6.2
Bar Chart of Total Sales by Company with Chart Components



Charts for structured data typically display the following components. Please take the time to learn these terms because they are essential for visualizing data.

- *Horizontal axis:* Used for labels. In Exhibit 6.2, the labels on the horizontal axis correspond to the individual companies, A–E.
- *Vertical axis:* Used for values. In Exhibit 6.2, the vertical axis starts at a value of $0 and goes up to a value of $40,000.
- *Scale:* The scale by which the values are measured. In Exhibit 6.2, the scale is 0 to 40,000.
- *Tick marks:* The increments of the scale that are listed on the vertical axis. In Exhibit 6.2, the tick marks appear in increments of 5,000, starting at zero: 0, 5,000, 10,000, 15,000, and so on. Tick marks should appear at equal intervals.
- *Data series:* The raw data the chart visually represents. In Exhibit 6.2, the raw data are total company sales.
- *Data point:* Each data series is made up of individual data points (values/categories). The bars in Exhibit 6.2 show the Total Sales data point for each company.

Sorting Considerations

In Exhibit 6.2, the Total Sales numbers are very similar, making it difficult to determine which column is the highest. We can therefore make the chart in Exhibit 6.2 more meaningful by sorting the data. (Charts in Excel default to sorting in the order of the data series presented in the original data set.)

To sort the data in Excel, ensure that the active cell is in the Total Sales data column, click the **Data** tab in the ribbon, and select **A-Z** to sort the numerical data in ascending order (Z-A for descending order) or **Data > Sort & Filter > Sort A to Z** (as shown in

Exhibit 6.3). If you want to sort by multiple columns, select **Data > Sort & Filter > Sort** and enter the columns you want to sort by (not shown in Exhibit 6.3).

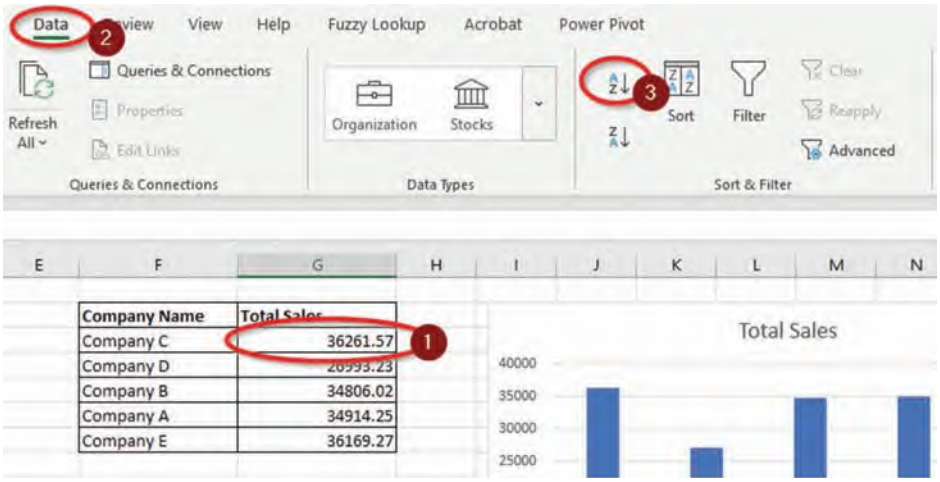


Exhibit 6.3
Sorting Data in Excel

As soon as the numerical data are sorted, the chart will shift, as shown in Exhibit 6.4. To make the bar chart even easier to read, we have added a descriptive title and X and Y-axis labels. In addition, the values for Total Sales on the Y-axis are formatted as currency to make it clear that Total Sales are measured in dollars.

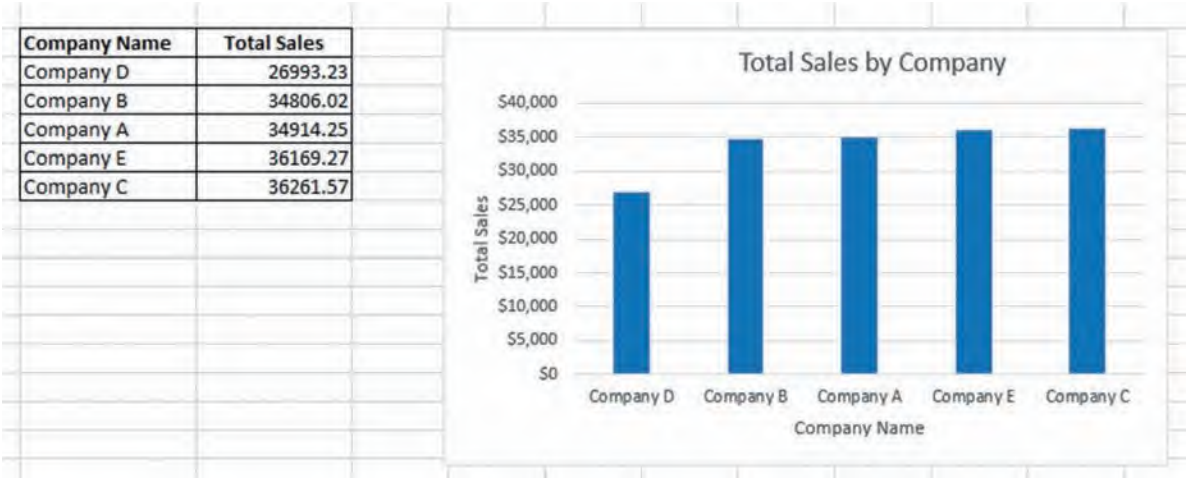


Exhibit 6.4
Bar Chart Sorted by Total Sales by Company

This sorted chart makes it easier to identify which company brought in the highest sales because the bars appear from left to right in ascending order. Thus Company D's bar appears first because it has the lowest sales, and Company C's bar appears last because it has the highest sales. Note that a chart with horizontal bars rather than vertical bars would also work well here.

In general, the type of data you are working with can help determine the attribute you choose to sort by. Recall these key terms from Chapter 2:

- **Categorical data** categorize items into groups that are represented either by words or by non-meaningful numerical data. For example, categorical data might group people by age (for example, adult, non-adult), transaction type (for example, sales versus returns), or specify an inventory costing method (for example, FIFO, LIFO, average cost).
- **Nominal data** are a type of categorical data that cannot be ranked, such as eye color and transaction type (purchase, return).
- **Ordinal data** are a type of categorical data with natural, ranked categories, such as letter grades (A, B, C, D, F) and Olympic medals (gold, silver, bronze).
- **Numerical data** are meaningful numbers that represent quantities, such as transaction amount, net income, age, or score on an exam.

In Exhibit 6.4, we are working with one numerical variable (Total Sales) and one categorical (nominal) variable (Company Name). When working with nominal data, it typically makes sense to sort based on the numerical value (Total Sales in Exhibit 6.4). However, if the categorical variable you are working with is ordinal, it may make sense to sort by the category instead of the numerical values, such as letter grade order starting with A.

PROGRESS CHECK

1. What is typically shown on the vertical axis in a data visualization?
2. What is typically shown on the horizontal axis in a data visualization?
3. How do you sort data in a chart in Excel?
4. Why did we sort the data in Exhibit 6.4 by the numerical value (Total Sales) instead of by the categorical value (Company Name)? Would it be useful to sort the data by Company Name instead? Explain.

LO 6.2

Distinguish among chart types.

6.2 DISTINGUISHING AMONG CHART TYPES

Determining the best method for visualizing your results requires you to answer four key questions:

1. What business question are you trying to answer?
2. Who is the target audience for the visualization?
3. Which type of data is being visualized, categorical or numerical?
4. What type of analysis have you performed, exploratory or confirmatory?

Recall from Chapters 4 and 5 the definitions of exploratory and confirmatory business analytics.

- *Exploratory business analytics* refers to the initial descriptive and diagnostic analytical investigations to summarize and explain performance. These types of analytics explore historic data and begin to generate questions that require further investigation and hypotheses that need testing. Exploratory business analytics often includes describing past performance, exploring summary performance statistics, identifying anomalies and outliers, checking assumptions, and detecting patterns.
- *Confirmatory business analytics* is conducted after the data analyst understands the relationships in the data. It uses statistics to judge the likelihood (probability) of a future event or outcome occurring. It includes predictive and prescriptive analytics.

Charts Appropriate for Categorical Data

Because categorical data and numerical (quantitative) data have different characteristics, they must be visualized with different types of charts. The charts most frequently used to visualize categorical data are:

- **pie charts**, which are circular graphs in which each slice represents the category's proportion of the whole
- **bar charts**, which are data visualizations that show categorical data using vertical/horizontal bars with varying heights/lengths proportional to the values they represent
- **stacked bar charts**, which are bar charts with the bars subdivided into different categories

The pie chart is probably the most common (and some would say infamous) data visualization for categorical data. It shows the parts of the whole; in other words, it represents the proportion of each category as it corresponds to the whole data set. A **proportion** is the number of observations in one category divided by the grand total of all observations.

A bar chart can also show the proportions of each category. In most cases, a bar chart is easier to interpret than a pie chart because our brains are more skilled at comparing the height of columns (or the lengths of horizontal bars) than they are at comparing the sizes of pie slices, especially when the proportions are relatively similar.

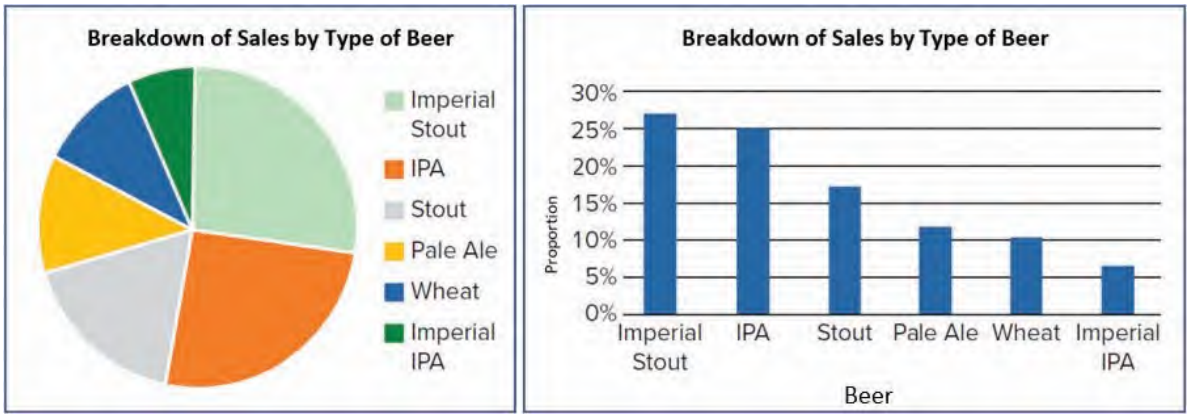


Exhibit 6.5
A Pie Chart and Bar Chart Visualize Proportion in Different Ways

Consider the two different charts in Exhibit 6.5. Both charts represent the proportion of beer type sold by a beer company. The size of the difference between the two largest pieces of pie (representing Imperial Stout and IPA) is almost impossible to see in the pie chart. This difference is easier to see in the bar chart. We could improve the pie chart by adding the percentages associated with each proportion, as shown in Exhibit 6.6, but it is much easier to see the difference in proportions by glancing at the order and length of the bars in the bar chart in Exhibit 6.5.

A good rule of thumb to follow with pie charts is to avoid using them when there are more than five categories. As the number of categories (slices of pie) increases, it becomes increasingly difficult to distinguish between the size of the slices. It is particularly difficult to read a pie chart with many skinny slices of pie.

We can also represent the same set of data in a stacked bar chart or a 100 percent stacked bar chart, as shown in Exhibit 6.7. A stacked bar chart can take less space on a page or a PowerPoint slide than a traditional bar chart. The chart on the left side of Exhibit 6.7 is a

Exhibit 6.6
Pie Chart Showing Proportion Detail

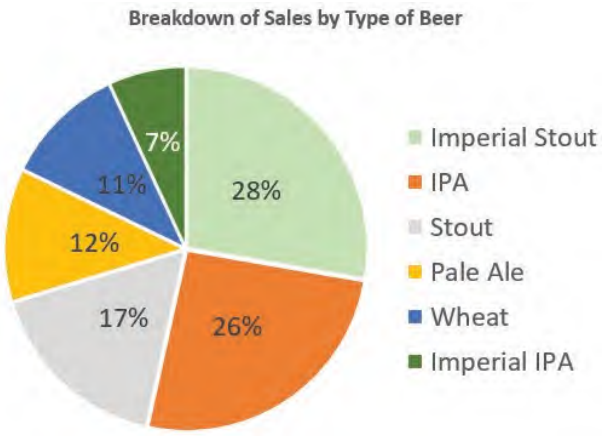
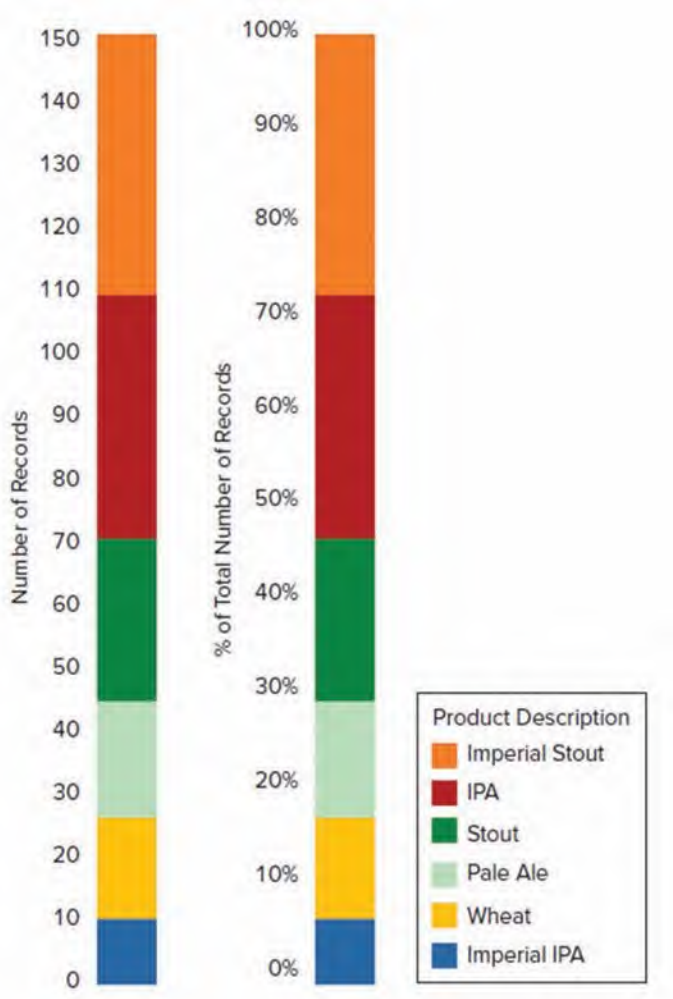


Exhibit 6.7
Stacked Bar Charts



stacked bar chart that shows the proportion of each type of beer sold as expressed by the number of beers sold for each product. The 100 percent stacked bar chart on the right side shows the proportion expressed in terms of percentage of the whole. Note that the Y-axes begin at zero and increase as the count (number) or percentage increases.

While it is sometimes difficult to compare two values in stacked bar charts, they work well when you are trying to break down and compare parts of the whole. For example, you could prepare a stacked bar chart with two columns, one for 2025 sales and one for 2026 sales, with each chart showing the breakdown by type of beer sold to visualize how the types of beer sold have changed over time. Similarly, if your company has multiple locations, you could prepare a stacked bar chart with columns broken down by product sold at each location to see relative product sales at different locations.

Several other types of charts also work well for showing proportions:

- *Tree maps.* A tree map utilizes size and color to show the proportional size of values using physical space. This type of visualization works well when there are many categories that need to be visualized. Tree maps are helpful in revealing patterns and identifying outliers. However, they are not ideal when you want to represent precise numbers or proportions.

  Exhibit 6.8 is a tree map of cases of beer sold. It shows that the two types of beer that sold the most cases are Imperial Stout and IPA. However, it is difficult to determine exactly how sales of Imperial Stout and IPA differ. Adding labels showing the number of cases sold for each beer would improve the readability of this tree map.

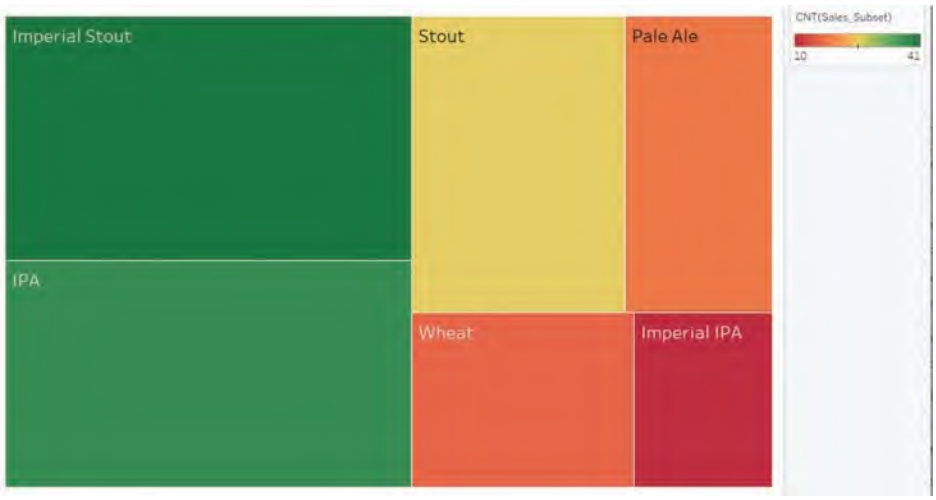


Exhibit 6.8
Tree Map (Cases Sold by Type of Beer)

- *Symbol maps.* A type of geographic map, a symbol map is useful for expressing categorical data proportions across geographic areas such as states or countries. Exhibit 6.9 is a symbol map that shows the number of sales by U.S. state. Notice that the symbols (in this case, circles) are smallest in Louisiana and Tennessee, indicating the lowest number of sales orders in those two states. Although a symbol map is useful in comparing relative proportions across different geographic regions, it is not the best choice for visualizing your data if a greater degree of precision is necessary.
- *Word clouds.* If you are working with text data, you can use a word cloud to identify the most commonly used words. The basis of a word cloud is the frequency of each word mentioned in a data set. The higher the frequency (proportion) of a given word, the larger and bolder the font for that word in the word cloud. Suppose you want to analyze the results of an open-ended response question on a survey. A word cloud is a great way to quickly identify the most commonly used words to determine if the respondents have a positive feeling or a negative feeling about what's being surveyed. When creating a word cloud, you can instruct the software to leave out the most commonly used English words—such as *the, an,* and *a*—so that important words are the focus of the analysis. Exhibit 6.10 is a word cloud for the text of Chapter 3 of this textbook.

Exhibit 6.9
Symbol Map (Sales by State)



Exhibit 6.10
Word Cloud (Based on Chapter 3 Text of This Textbook)



- *Heat maps.* A heat map is a graphical representation of categorical data that uses different colors to indicate different values. Different colors, or different shades of the same color, allow users to compare the intensity of a phenomenon. Exhibit 6.11 is a heat map for weekday unit sales by employee. The darker the red color, the fewer units an employee sold. The darker the blue color, the more units an employee sold. Based on the heat map, we can see that Kobe is selling the fewest units, while Mykalya is consistently performing well.

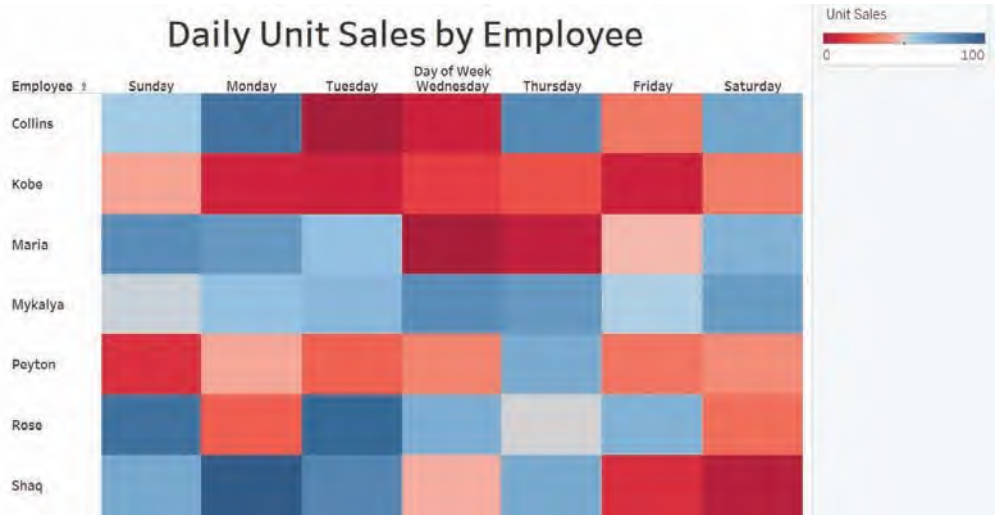


Exhibit 6.11
Heat Map (Daily Sales by Employee)

Charts Appropriate for Numerical (Quantitative) Data

Before we discuss visualizations for quantitative data, let's review some definitions from Chapter 3 about numerical data, which represent meaningful quantities, such as transaction amount, net income, age, or exam score.

- **Continuous data** are numerical data that can take on any numerical value, not just whole numbers, meaning there is an infinite set of values between any two observations in a continuous data scale. Examples of continuous data are height, weight, and currency.
- **Discrete data** are numerical data that only take on whole-number (integer) values and have a finite set of values between any two observations. An example of discrete data is the number of products in inventory (you can have 0 items or 150 items in inventory, but you cannot have 147.8 items in inventory).

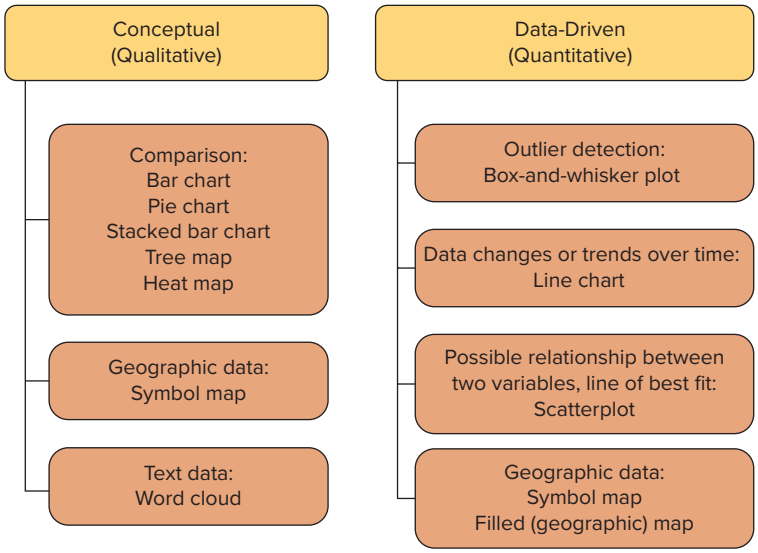
There are many methods for visualizing quantitative data. Except for the word cloud, all the methods mentioned in the previous section for visualizing categorical data can work for depicting numerical data, but the following chart types are particularly useful for depicting more complex data. We discuss them in more detail, and provide examples of each, a bit later in this chapter.

- *Box-and-whisker plots and histograms.* As we saw in Chapter 3, these are useful for visualizing the distribution of a data set and identifying outliers.
- *Line charts.* These charts are well suited to showing data changes or trend lines over time. Line charts are useful for visualizing continuous data, while bar charts are often used for visualizing discrete data. Line charts are not recommended for qualitative data, which are categorical and therefore can never be continuous.
- *Scatterplots.* A **scatterplot** is a graph that plots two variables each on its own axis. Scatterplots are useful for identifying the possible correlation between two variables and for identifying a trend line or line of best fit.
- *Filled (geographic) maps.* Unlike a symbol map, a filled geographic map uses shading, tinting, or patterns to display how a value differs across geographic regions, such as states or countries.

Exhibit 6.12 provides a summary of best practices for the chart types just described. Each chart option works equally well for exploratory and confirmatory data visualizations. For example, when you want to compare numerical variables, a bar chart is an optimal

choice, but the other choices shown (pie chart, stacked bar chart, tree map, heat map) might work just as well or better. However, this exhibit shouldn't stifle your creativity; many of the charts in the listed categories can work well with other data types, depending on your purpose.

Exhibit 6.12
Summary of Chart Types



Like selecting and refining your analytical model, reporting results is more art than science. Your goal should always be to share critical information with stakeholders clearly and concisely. Although visualizations often make a strong impact, they can become a distraction if you're not careful. Worse, they can misrepresent data. For example, bar charts can give the wrong impression if the Y-axis starts above zero or is divided into very large or very small increments. Similarly, 3D graphs can be extremely deceptive because they may distort the scale even if the numbers are accurate.

PROGRESS CHECK

5. What type of chart would you use to summarize the results of a text analysis of **Home Depot**'s annual financial report?
6. Why must we be careful when drawing conclusions from data visualization?

LO 6.3

Describe appropriate visualizations for exploratory business analytics.

6.3 VISUALIZING EXPLORATORY BUSINESS ANALYTICS

Recall from Chapter 4 that **exploratory business analytics** refers to the initial descriptive and diagnostic analytical investigations to summarize and explain performance. You can use data visualization to explore historic data and begin to generate questions and hypotheses that need further explanation.

Bar Charts, Line Charts, and Pie Charts

Bar charts, line charts, and pie charts can all be useful in exploratory business analytics. Let's begin with a simple example. Exhibit 6.13 is a simple table showing **Amazon**'s net income and sales from 2012 through 2021. Presented in tabular form, it is difficult to see how net income and sales change over time.

Year	Net Income	Sales
2012	(39)	61,093
2013	274	74,452
2014	(241)	88,988
2015	596	107,006
2016	2,371	135,987
2017	3,033	177,866
2018	10,073	232,887
2019	11,588	280,522
2020	21,331	386,064
2021	33,364	469,822

Exhibit 6.13
Table of Financial Performance for Amazon, 2012–2021 ($ in millions)

Source: Amazon Income Statements 2012–2021

When the same data are visualized using a bar chart, as in Exhibit 6.14, you can quickly see that Amazon's sales have grown steadily each year. Note the chart components and how they align with the basic elements of a chart that you learned in earlier in this chapter:

- The vertical axis has a scale from 0 to 500,000, with tick marks indicating each incremental 50,000 increase.
- The labels on the horizontal axis indicate the years from 2012 through 2021.
- The bars indicate the data series with the total sales data point for each year.

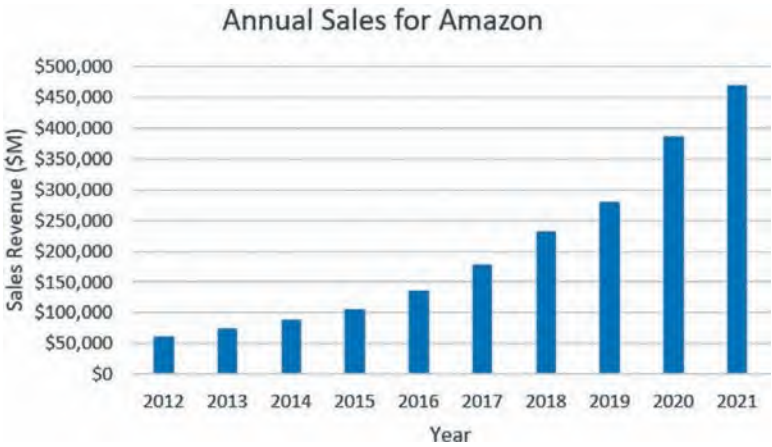


Exhibit 6.14

Source: Amazon Income Statements 2012–2021

Bar charts should be sorted logically. When visualizing **time series data** (values taken on by a variable at different points in time: days, months, or years), it typically makes sense to sort the data chronologically. However, bar charts also lend themselves to sorting based on the numerical value of the variable. In this case, Amazon's sales increased each year, so the bar graph would not change if we sorted it in ascending order based on sales.

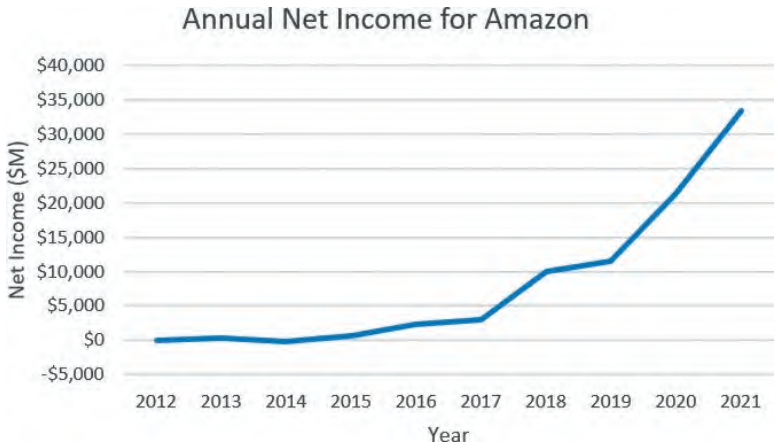
Because Amazon's sales data are time series data, we can also use a line chart to visualize the data effectively. Line charts should always be sorted chronologically. Recall the bar chart in Exhibit 6.4, which compared sales from five different companies. Using the same set of data to create a line chart would not make sense because there is nothing that

inherently connects the different companies' sales totals. However, if you had yearly sales data for each of those five companies, you could place those data in a line chart, with one line per company, to see how total sales compare across the companies over time.

The data for **Amazon**'s net income in Exhibit 6.13 can be depicted as a line chart (Exhibit 6.15) because the net income is provided over a 10-year period. In addition to being useful for visualizing time series data, line charts work well in visualizing numerical data that extend below 0. In this example, Amazon had negative income in 2012 and 2014.

Exhibit 6.15
Line Chart of Amazon's Net Income, 2012–2021 ($ in millions)

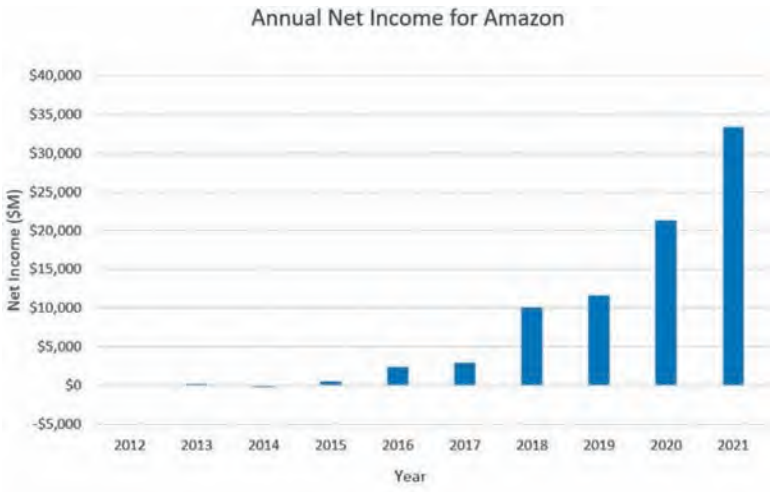
Source: Amazon Income Statements 2012–2021



These net income data could also be depicted with a bar chart, as shown in Exhibit 6.16. However, this bar chart is clunkier than the line chart. The data point for 2012 is barely visible on the bar chart because it is very close to 0, and the negative value for 2014 is also difficult to see. Line charts are preferable when you want to communicate the overall trend rather than specific data points. In addition, line charts provide more flexibility with scales.

Exhibit 6.16
Bar Chart of Amazon's Net Income, 2012–2021 ($ in millions)

Source: Amazon Income Statements 2012–2021



Typically, ratio data should be visualized with 0 as the starting point, especially in a bar chart. Recall that **ratio data** are numerical data with an equal and definitive value between each data point and a meaningful zero. An example of ratio data is net income. Unlike some ratio data examples where below-zero does not exist (for example, negative pounds or negative age), net income can be less than zero, making it necessary to sometimes start the chart below zero, as in Exhibit 6.16. In a line chart, it may make sense for the starting point to be a value above or below zero depending on the trend.

Pie charts are also useful for showing proportion and visualizing categorical data as a part of exploratory business analytics. Recall that categorical data tend to be represented by words, such as transaction types (for example, sales and returns). Pie charts show proportion, which is calculated as the number of observations in one category divided by the grand total of observations. However, note that Exhibit 6.17 is a poor use of a pie chart. The proportion of each year's sales over a 10-year period is not particularly useful; dollar values are included, but percentages are not. Also, it is difficult to determine the year that each slice represents, even though a legend is included at the bottom. As noted earlier, pie charts are rarely preferred when there are six or more categories (or "slices" of pie).



Exhibit 6.17
Pie Chart of Amazon's Sales, 2012–2021 ($ in millions)

Source: Amazon Income Statements 2012–2021

PROGRESS CHECK

7. When is a line chart preferred over a bar chart?
8. When visualizing data over time, which charts are the best options?
9. What type of data does a pie chart present?

Bar Charts Versus Histograms

Recall from Chapter 3 that **histograms** illustrate frequency in the distribution of data. Decision-makers can use histograms to assess the distribution of a particular data set and to compare the number of observations across intervals.

Although histograms look like bar charts, these two types of visualizations have three key differences.

1. *Histograms represent bins/intervals.* The bars of a histogram do not represent categories (such as product name or year). Instead, they represent bins. **Bins**, or **class intervals**, are subsets of the data, arranged in increasing order. The first bin begins with the lowest number in the range of data, and the last bin ends with the largest number in the range of data. Tableau automatically creates bins for you, but depending on the version of Excel you use, you may have to create your own bins. Microsoft Power BI does not offer a built-in option for creating histograms.

   Each bin must be the same size, and the bins must contain the entire range of data. Typically, a histogram has between 5 and 20 bins, depending on the size and diversity of the data set. It is a best practice to make the data set evenly divisible by the number of bins.

2. *Histograms use numerical data.* While bar charts can represent counts of categorical data, histograms are used only for numerical data. There are no gaps between the bars in histograms, unless the gap represents the absence of data in a particular bin.
3. *The vertical axis of a histogram shows count of observations.* While the vertical axis in a bar chart can be associated with a variety of descriptive statistics, the vertical axis in a histogram is always associated with the count of observations in each bin. Recall that **descriptive statistics** are summary statistics that briefly describe or summarize features of a data set or variable. The basic descriptive statistics include the mean, standard deviation, minimum, and maximum of a data set.

Let's assume that a bank wants to answer the following question: How long do customers wait for service? To answer this question, the bank collects data on customer wait time for twenty random observations. The results are shown in Exhibit 6.18, which shows that most customers appear to be served quickly, between 35 and 40 seconds. We know this exhibit is a histogram because (1) the bars are contiguous bins representing intervals of time, which are numerical data, and (2) the Y-axis is the frequency or count of observations.

Exhibit 6.18
Histogram for Customer Wait Time



To better understand the difference between a bar chart and a histogram, look at Exhibit 6.19, which answers this question: What is the value of our past-due accounts receivable? *Accounts receivable* refers to the amount of money owed to a company by its customers. When a company accepts payments on credit with an accounts receivable account, it needs to know when (and if) the accounts receivable balance has been paid. *Accounts receivable aging* is a popular technique for determining if credit sales are ultimately going to be collected. The rationale is this: The longer the receivable is due and outstanding, the lower the likelihood that it will ever be collected.

To perform this analysis, analysts can bin and visualize the number of days that accounts are late. They can place a company's receivables in 30-day buckets to show the total receivable dollars that are due and outstanding for 1–30 days, 31–60 days, 61–90 days, 91–120 days, 121–150 days, 151–180 days, and so on, as in Exhibit 6.19, which was created from an Excel pivot table.

We know that Exhibit 6.19 is a bar chart and not a histogram for the following reasons:

- The vertical axis represents the sum of the transaction amounts in each bin, instead of a count of transactions.

- The first bin starts at 1 (day), instead of the lowest value in the range of data.
- There are spaces between the bars.

Nonetheless, analysts will interpret this chart in much the same way they will interpret the results of a histogram. There does not seem to be a significant skew, with the bulk of the receivables in the middle intervals of the data set. However, the analysts should investigate the receivables over 90 days past due to understand why so many receivables may be uncollectible.

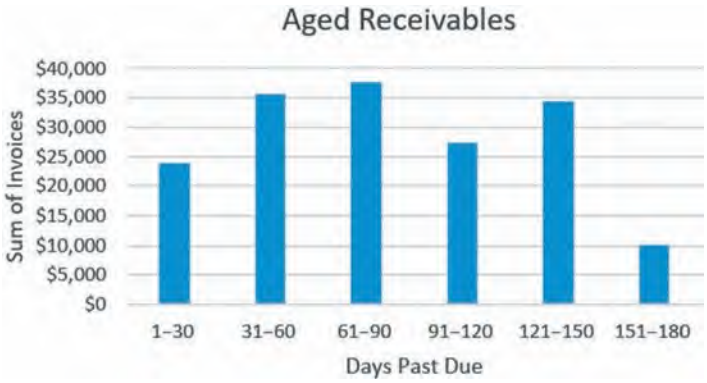


Exhibit 6.19
Aged Receivables from an Excel Pivot Chart

For the sake of comparison, look at a count of the aged receivables transactions in the histogram in Exhibit 6.20. The X-axis has the same label as the bar graph, but the Y-axis in the histogram is the frequency or number of observations in a bin as opposed to the dollar value of the invoices (see again Exhibit 6.19).

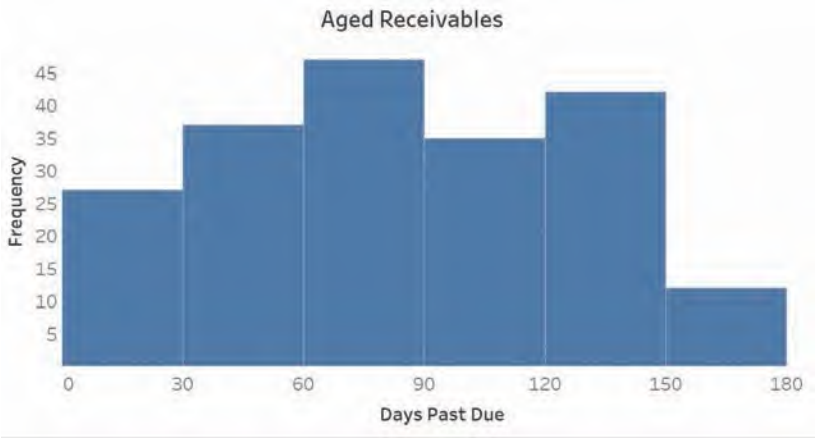


Exhibit 6.20
Histogram of Aged Receivables

PROGRESS CHECK

10. How do bar charts and histograms differ?
11. What is represented on the vertical axis in a histogram?
12. Draw an example of a positively skewed (or right-skewed) histogram. If exam results are positively skewed, what does that imply about the exam? Explain. (If necessary, review the definition of skew in Chapter 3).

Visualizing Outliers and Anomalies

Recall that diagnostic analytics addresses the following questions: "Why did it happen?" and "What are causes of past results?" and "Why are the results different than expectations?" Outliers, or extreme values in the data set, are often used as part of diagnostic analytics and can be assessed in two ways:

- "Eyeball" the data to determine if one or more observations look like they fall outside of the expected distribution.
- Construct a box plot to assess for outliers.

Let's begin with the simple method of eyeballing the data.

MINI CASE: "Eyeballing" Class Exam Outliers

A professor gave an exam to a business analytics class. Exhibit 6.21 presents the grade distribution via a histogram, showing how many As, Bs, Cs, Ds, and Fs were earned on the test. We can tell that the exam whose scores are summarized in Exhibit 6.21 was relatively easy because of the shape of the distribution in the histogram. As explained in Chapter 3, this distribution is *negatively skewed* or *skewed left* because the bulk of the data is in the upper range of the distribution (Cs or 70–79, Bs or 80–89, and As or 90–100), but there is an outlier to the far left, or the lower end of the distribution.

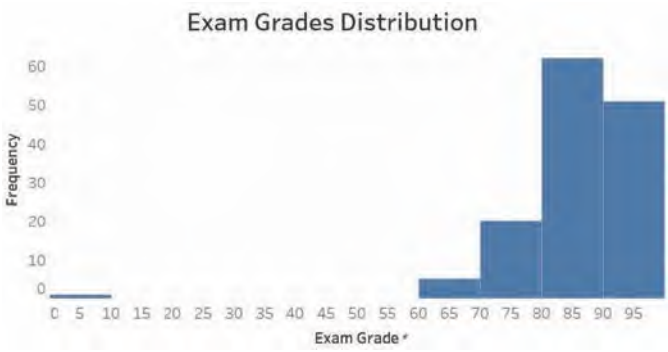


Exhibit 6.21
Histogram of Exam Grade Distribution

Decision-Making with Data

Which scores should be investigated? What are potential reasons for the outlier score? For example: Did a student fail to show up for the exam? Did one student just perform poorly on this exam? It's important to consider all possibilities; sometimes an outlier represents an error in the data and therefore needs to be corrected. Other times, the outlier doesn't represent an error but rather an observation that needs to be investigated.

The second method of identifying outliers, constructing a box-and-whisker plot (or box plot), is more technical. Recall from Chapter 3 that box plots are similar to histograms in that both represent the distribution of a data set. Unlike histograms, however, box plots pack a lot of descriptive statistics into one graph including mean, median, percentile, outliers, and expected minimum and maximum.[3]

[3]The maximum expected value is equal to 75th percentile + 1.5 * the interquartile range = 1,154,14 + 1.5 * (1,154.14 – 750.77) = 1,759.20. The minimum expected value is equal to 25th percentile – 1.5 * the interquartile range = 750.77 – 1.5 * (1,154.14 – 750.77) = 145.72.

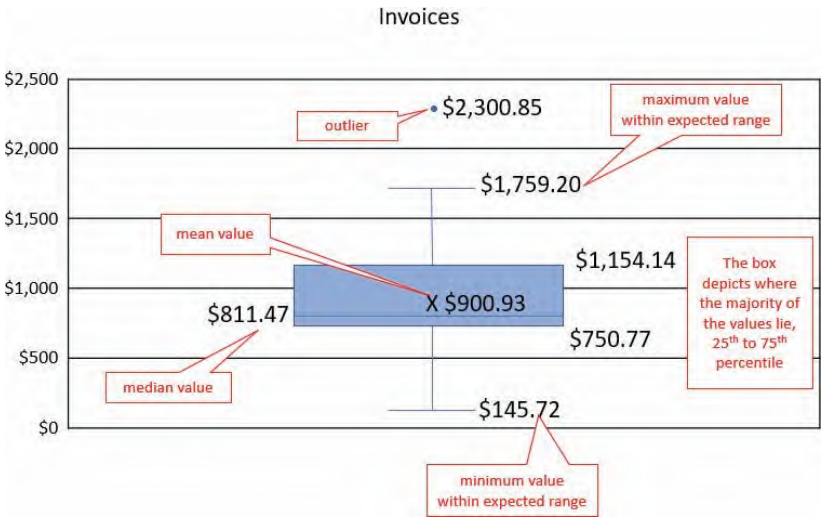


Exhibit 6.22
Box-and-Whisker Plot for Invoice Amounts

Because the majority of the data lies between $750.77 and $1,154.14 and the expected values lie between $145.72 and $1,759.20, it seems reasonable to investigate the outlying value, $2,300.85, to see if it represents an error. This visualization makes it easy to identify extreme values.

The utility of box plots does not end with the identification of potential outliers. Assessing the range of the plotted variable, the dispersion of variables, and the skewness (if any) is helpful even in the absence of outliers.

Pivot Tables

In Excel, one of the best methods for exploratory analytics is pivot tables accompanied by pivot charts. Data in a pivot table can be sliced, diced, and filtered to provide additional insight about a data set. You can also visualize your data using conditional formatting or sparklines (both introduced in Chapter 4).

Let's consider an example in which a pivot table with an accompanying pivot chart can yield additional insight. A fictitious company, Fruity Inc., sells 79 different fruits. The company wants to answer the question: Which fruit has the largest total gross margin for all sales in March? Gross margin for a sale equals the sales price minus the cost of goods sold. Exhibit 6.23 presents an Excel pivot table and corresponding pivot chart (here, a bar chart) for the top five fruits in terms of total gross margin. If you click on the Apple bar, Excel drills down to the detail of the sales making up the bar, as shown in Exhibit 6.24.

Fruit	Sum of Gross Margin
Apple	140.39
Orange	110.65
GrapeRaisin	89.29
Soursop	79.92
Apricot	78.02
Grand Total	**498.27**

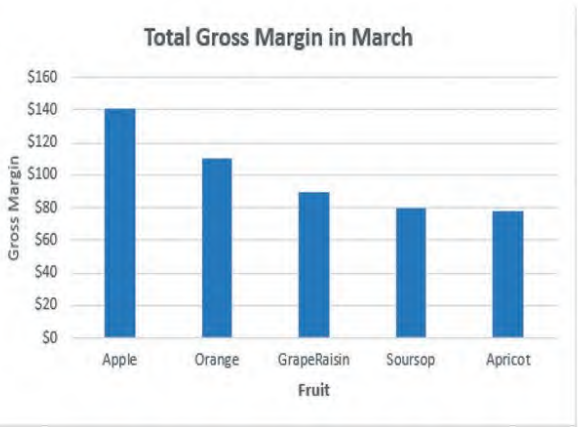


Exhibit 6.23
Example of Excel Pivot Table and Pivot Chart

Invoice #	Customer #	Zip Code	Date of Sale	SKU	Description	Quantity (Pounds)	Sales	Cost	Gross Margin
3021	2001	67510	3/1/2026	FRT-APL	Apple	15	25.52	16.36	9.16
3069	2047	67838	3/4/2026	FRT-APL	Apple	20	34.02	21.81	12.21
3067	2072	66736	3/4/2026	FRT-APL	Apple	50	85.02	54.51	30.51
3061	2024	66937	3/3/2026	FRT-APL	Apple	15	25.52	16.36	9.16
3056	2063	67335	3/3/2026	FRT-APL	Apple	10	17.02	10.91	6.11
3050	2084	66936	3/2/2026	FRT-APL	Apple	10	17.02	10.91	6.11
3043	2087	67731	3/2/2026	FRT-APL	Apple	50	85.02	54.51	30.51
3040	2074	67422	3/2/2026	FRT-APL	Apple	15	25.52	16.36	9.16
3030	2063	67010	3/1/2026	FRT-APL	Apple	45	76.52	49.06	27.46

Exhibit 6.24
Example of Excel Drill Down for Apple Sales

LAB CONNECTION

LAB 6.1 provides step-by-step instructions for performing exploratory analytics in Excel. Specifically, it asks you to use conditional formatting and sparklines to quickly view trends.

PROGRESS CHECK

13. In Exhibit 6.21, how many students earned a C, assuming an A = 90 or above, B = 80–89, and C = 70–79?
14. In Exhibit 6.21, how many students earned above a C, assuming an A = 90 or above, B = 80–89, and C = 70–79?
15. Refer to Exhibit 6.22. Which statistic is represented by an "X" in the box plot? Which statistic is represented by a horizontal line in the box plot?
16. Identify at least two observations you can make about a set of data when you are analyzing a box plot.
17. Why is creating a box plot more helpful than just eyeballing the data?

LO 6.4

Describe appropriate visualizations for confirmatory business analytics.

6.4 VISUALIZING CONFIRMATORY ANALYTICS

Recall that **confirmatory analytics** refers to predictive and prescriptive analytics that use statistics to judge the likelihood (probability) of a future event or outcome occurring. Whereas exploratory business analytics considers generally known facts, confirmatory business analytics provides less deterministic output and more probabilistic models.

In addition, confirmatory analytics should tell a story. Often, telling this story requires adjusting the defaults in the software tool and highlighting a particular data point, bar, or line. To determine if your data visualization tells a story, take a good look at your visualization and determine whether your audience will know the precise question you are answering. If they will not, then you may need to change the format or the chart type.

To begin our discussion of confirmatory analytics, we discuss predictive analytics, which is performed by identifying patterns in historical data to judge future likelihood or probability. Correlation, regression, and forecasting are three common methods used in confirmatory analytics. The results of all of these methods can be visualized to provide greater insight to decision-makers.

Correlation and Regression

Predictive analytics can be particularly useful when there is a strong correlation between the past and future. **Correlation** is a measure that helps us understand if variables are related to each other and if so, to what extent. (For a review of correlation, see Chapter 3.) It helps us understand if two variables change in the same direction, different directions, or not at all. Correlation values can range from –1 to 1, and they are often expressed as a percentage. For a correlation between two variables to be considered high (that is, there is a strong relationship between the two variables), the correlation value is closer to –1 or 1. A negative correlation value (close to –1) indicates that an increase in one variable is associated with a decrease in the other variable, while a positive correlation value (close to 1) indicates the variables move together in the same direction. For the correlation to be considered low (that is, there is very little relationship between the two variables), the correlation value will be close to 0.

Correlations, which can be calculated in Excel, can be enhanced with a visualization such as a scatterplot. Recall that scatterplots plot the values of two variables, each on its own axis. If the variables are correlated, then the variables are related in a linear fashion and the points in the scatterplot resemble a line pattern (without an actual line being drawn). For example, Exhibit 6.25 shows the first several rows of a company's sales amounts across different states for 2025 and 2026. (The full data set contains many more rows; we will see them visualized in Exhibit 6.26.)

State	2025	2026
Alabama	7,651	1,828
Arizona	6,242	11,134
Arkansas	2,224	2,708
Cailfornia	131,552	146,388
Colorado	10,667	10,300
Connecticut	3,998	5,307
Delaware	2,720	13,755

Exhibit 6.25
Sales Data for 2025 and 2026

Simply looking at the values doesn't help us determine if there is a correlation between the two years for each state. However, placing the data into a scatterplot (Exhibit 6.26) can

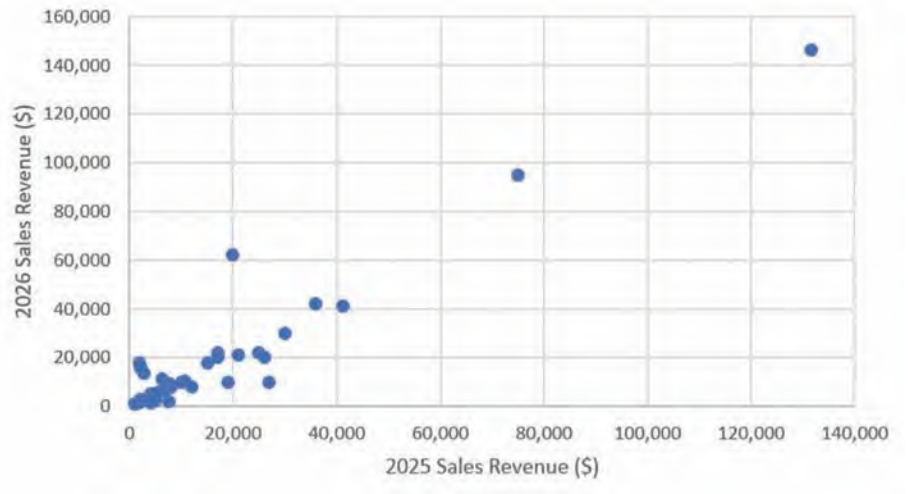


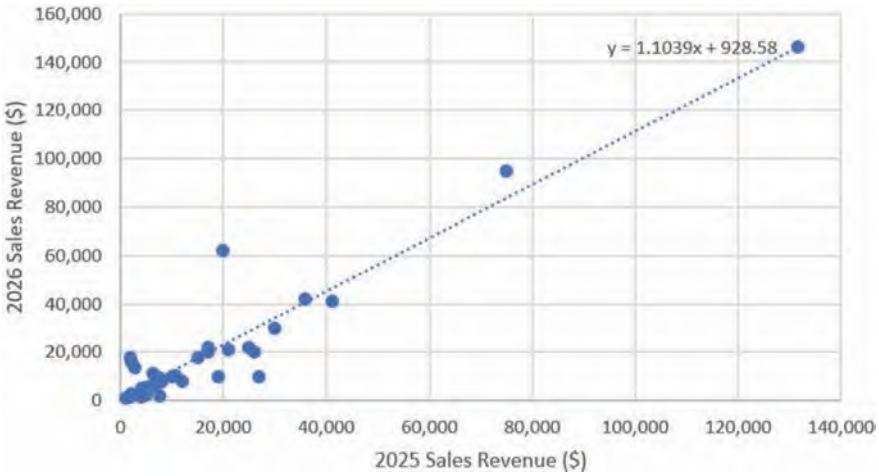
Exhibit 6.26
Scatterplot of Correlation Between Two Variables

help us visualize the relationship between sales in the two different years. The data points in this scatterplot reveal a tendency: When a value in *X* increases (2025 sales), the value in *Y* (2026 sales) also increases. There is a correlation between the two variables, and it is positive, as depicted by the data points (which could be connected by a line) rising upward to the right.

When two variables, such as past sales and present sales, are correlated, we can use the past as a valuable predictor of future events. To create a model to predict the future based on the past, you can run a **regression** to get more specific information about the relationship between inputs and outputs. (For more detail, see Chapters 3 and 5.) After creating a scatterplot, you can plot a trend line to visualize a linear relationship. A **trend** shows the general direction in which a variable is moving. The trend line will help visualize how well the linear regression model "fits" the actual observations, or data points. Trend lines can also help identify outliers that are pulling the trend in a drastic direction. Exhibit 6.27 depicts the same scatterplot from Exhibit 6.26, but with a trend line and the corresponding regression equation.

Exhibit 6.27
Regression Analysis: Scatterplot with Trend Line and Line of Best Fit

Tableau Software



Forecasting with Time Series Data

Not only can you visualize events that have already happened (such as sales), but you can also predict the future with a forecasting line chart to show possible scenarios. Exhibit 6.28 illustrates quarterly sales over time, using blue to indicate current sales and orange to indicate three predicted scenarios—**forecast** (most likely sales outcome), lower confidence (worst-case level of sales), and upper confidence (highest predicted level of sales).

The sales of many products exhibit seasonality, which occurs when sales tend to rise or fall at the same time each year. For example, sales of snow shovels and rock salt rise in the winter and fall in the summer. Sales of beachwear rise in the summer and fall in the winter. In Exhibit 6.28, quarterly sales drop in the first and second quarters of each year. The forecast captures this seasonality.

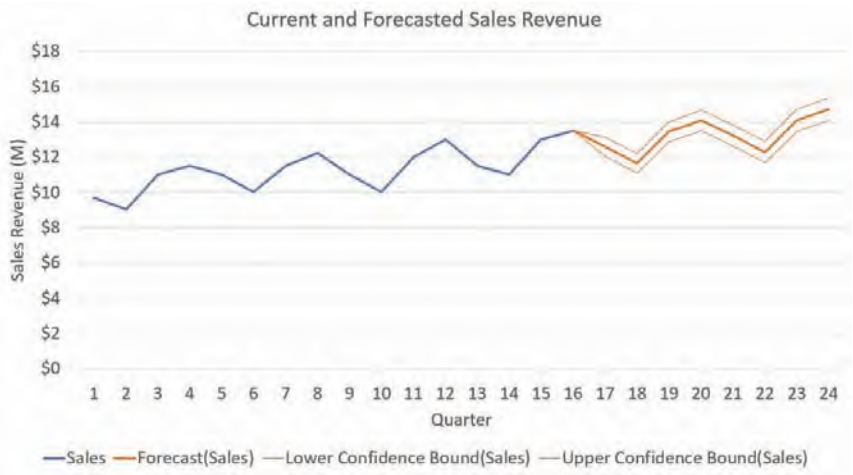


Exhibit 6.28
Time Series and Forecast Visualizations

PROGRESS CHECK

18. Which two phenomena in the real world are positively correlated? (There are many possible correct answers.)
19. Which two phenomena in the real world are negatively correlated? (There are many possible correct answers.)
20. What does it mean when the correlation between two variables has a value of 0?

6.5 PRESENTING DATA IN A DASHBOARD

LO 6.5

Explain dashboards and their benefits.

When you think of a dashboard, you probably think of your car, whose dashboard presents all the information you need in order to make driving decisions: a speedometer to ensure you're keeping under the speed limit, an odometer to indicate when to take your car for service (or to think about buying a new car), a gas gauge to help prevent you from running out of fuel, and so on. In a way, your car's dashboard tells your car's life story.

Similarly, in a business setting, a dashboard tells the business's life story. It presents a collection of useful visualizations and tables to check business progress over a period of time or at a particular point in time. The dashboard should include measures that companies use to make business decisions, such as net income, total sales revenue, cost of goods sold, number of customers, and sales by product. While many measures can be included in a dashboard, dashboards should focus on the key performance measures essential to analyzing performance and making decisions. Otherwise, the dashboard may cause information overload, which may hinder rather than help decision making. It is also helpful if dashboards are interactive, allowing users to dig into the data by filtering one or more of visualizations. For example, a marketing dashboard may allow users to isolate an individual customer or group of like-aged customers to reveal total sales revenue for the individual/group as well as sales by product.

MINI CASE: A Sample Excel Dashboard for Amazon

Earlier in this chapter, we looked at **Amazon**'s net income and sales in tabular form and learned how to transform those data into a data visualization. We can make that analysis even more meaningful by displaying both the year-over-year sales bar chart and the year-over-year net income line chart in the same place—that is, in a dashboard. Dashboards typically have filters available so that decision-makers can easily hone in on a particular time period, category, product, or geographic area. If we had more detailed information about the different locations and products that produce the aggregated data in Exhibits 6.14 and 6.15, we could filter the two charts so that they shift based on the different location, product category, or year we selected, as shown in Exhibit 6.29.



Exhibit 6.29
How Amazon May Use a Simple Dashboard

Decision-Making with Data

Examine the dashboard in Exhibit 6.29. Which additional data elements would be helpful to decision-makers at Amazon? What additional style elements (for example, colors, organization, placement) would improve the dashboard? Is it helpful to have both graphs and a table of net income and sales? Which is more useful: the line graph or the bar chart? Explain.

It is important to remember that a good dashboard consists of key performance measures as well as effective individual data visualizations that work together to tell the story of the company. The individual data visualizations should each communicate the results of analytics to the intended audience using a variety of charts, including graphs, plots, and maps. The next section provides a checklist that will help you create impactful charts.

Checklist for Creating Effective Charts That Clearly Answer Business Questions



In *How Charts Lie,* Alberto Cairo describes the many ways that a chart might lie, confuse, manipulate, and mislead.[4] Culprits include:

- poor design
- the use of incorrect or an inappropriate amount of data
- the concealment of data
- the suggestion of misleading patterns
- support for pre-existing desired outcomes, opinions, or assumptions
- unclear communication of uncertainty.

For example, the cone of uncertainty used by the U.S. National Hurricane Center (USNC) indicates the likelihood of a hurricane path. The less certain meteorologists are of the hurricane's path, the wider the cone. However, most people incorrectly interpret the map as depicting how the hurricane will strengthen over time, which means that the USNC's chart is problematic.

To become a skilled and ethical creator of data visualizations, Cairo recommends examining many charts and graphs while "being aware of . . . biases and learning to see what charts and graph *don't* necessarily show." [5] He also recommends applying the Golden Rule when you are creating charts: If you don't like being tricked, don't trick other people.

As you think about the data visualizations that you will use to convey information, consider the following questions: How often do you think people deliberately use charts and other data visualizations to mislead or deceive others? What questions should you ask about charts created by others? What questions should you ask about your own charts before you make them public? How might your chart be misinterpreted? Why should you look at the data underlying a chart?

The following best practices will help you create effective charts, whether they stand alone or appear on a dashboard.

- Pick the chart type that is most effective in clearly communicating the results at the desired precision level.
- In general, answer one business question per chart. (Exceptions are possible.)
- Make the chart easy to read and interpret by always including a chart title, axis labels, tick marks, and data labels.
- Include a legend when more than one color or line type is used.
- Pick a scale with a beginning point that is not misleading and that uses natural, consistent intervals.
- Include all relevant data. The data you select should not skew results or mislead users. The breakdown of the data by time (minutes, hours, days, weeks, months, quarters,

[4]This section is based on Tricia Bisoux, December 12, 2019, "The Ethics of Data Visualization." Available online at https://www.aacsb.edu/insights/articles/2019/12/the-ethics-of-data-visualization. Accessed 04/22/2022.

[5]Tricia Bisoux, December 12, 2019, "The Ethics of Data Visualization." Available online at https://www.aacsb.edu/insights/articles/2019/12/the-ethics-of-data-visualization. Accessed 04/22/2022.

years) depends on the question you are answering, the data, the analysis, the results, and the audience's use of the data.
- Use different colors to differentiate different data types and highlight focal points, not to make the data visualization "pretty."
- Use bright colors for emphasis, but use them sparingly because the human brain interprets too many bright colors as "nothing is highlighted."[6]
- Use colors in accordance with generally accepted interpretations. For example, red is associated with stop and negative profits, while green is associated with go and positive results. Therefore, do not use red to represent positive company profits, which will confuse users.
- Include enough relevant information to inform your audience without overwhelming them.

LAB CONNECTIONS

LABS 6.2, **6.3**, **6.4**, and **6.5** provide step-by-step instructions on how to create different types of dashboards.

- **LAB 6.2** creates a static dashboard to present information in a professional report format in Tableau and Power BI.
- **LAB 6.3** creates an interactive dashboard in Tableau and Power BI using filters.
- **LAB 6.4** creates an interactive dashboard in Excel using pivot tables and slicers.
- **LAB 6.5** creates an interactive dashboard in Tableau and Power BI.

PROGRESS CHECK

21. What do we mean when we say that a dashboard is interactive?
22. Why does a dashboard support decision-making better than a single data visualization?

LO 6.6

Explain how to communicate your analysis results in a written summary or report.

6.6 COMMUNICATING YOUR DATA WITH WORDS: EXECUTIVE SUMMARIES AND REPORTS

After you have analyzed your data, it is often useful to summarize your findings in words, not just in data visualizations. Depending on the amount of information you want to convey and how much your audience is interested in reading your analysis, you may choose to prepare an *executive summary* or a *full report* (which usually includes the executive summary).

Executive Summary

An executive summary, which is typically less than a page long, includes the following key sections:

- *Statement of the problem, question,* or opportunity that initiated the business analytics project.

[6]For more in-depth information, see Anneli Joplin, November 8, 2017, "Strategies for Effective Data Visualization." Available online at https://vdocuments.net/strategies-for-effective-data-visualization-trends-more-effectively-make-emphasis.html?page=5. Accessed 02/02/2022.

- *A concise description of why the project is relevant to the audience.* It is important to tailor all communications (written and otherwise) to your intended audience. Be careful not to include jargon, acronyms, or statistical terms that your readers may not understand.
- *Summary of the results.*
- *Brief recommendation(s)* for a course of action based on the results.

Exhibit 6.30 provides an example of an executive summary for Exhibit 6.18 (bank customer wait times).

Full Report

If you are providing a full report, the executive summary should serve as an introduction to and summary of what your audience will learn throughout the rest of the report. In addition to an executive summary, a full report should include:

- *A description of the method* chosen for the analytics project.
- *Expanded summary of the results discussed in the executive summary.* The report should go beyond the brief results provided in the executive summary to include:
  - *Pertinent data outputs,* such as correlations, results from tests of significance, and results from regression tests. When providing these data in a report, it is necessary to interpret the results. For example, do not simply report a p-value of 0.037. Instead, explain what that value means.
  - *Data visualizations that help to communicate the findings.* Provide a written summary of the information that the visualization provides. For example, if you provide a line chart that is skewed positively or negatively, explain what the skew signifies.
- *Expanded recommendation(s)* from the executive summary that reiterate the significance of the project and how it will affect your audience.

Regardless of the report type selected, successful communication requires effective storytelling that describes the data accurately and without bias. To achieve that goal, charts must honestly report the results using the proper chart type, accurate labels, all of the relevant data, and appropriately scaled axes with an appropriate starting point.

Effective storytelling also requires understanding your target audience. The target audience determines whether you should create an executive summary or a full report. The target audience also determines the appropriate format and language of the written report as well as types of charts included in the report. For example, a production line manager will more likely want to see statistical charts than a CEO will.



Data Visualization and Your Career

A recent search of **indeed.com** for job postings containing the phrase "data visualization" yielded over 72,000 job listings! One potential career for people with a strong understanding of business analytics is a business intelligence analyst, who analyzes operational data to monitor product usage, creates forecasts, and helps answer important business questions. Another potential career is a data analyst, who analyzes data across the company and creates visualizations to uncover hidden trends and insights. Both types of analysts create reports, data visualizations, and dashboards to share their findings. They need strong verbal and written communication skills, programming knowledge, and an analytical mindset. The average annual salary for careers in data visualization is $60,000 to $100,000.[7]

[7]Priya Pedamkar, 2020, "Careers in Data Visualization." EDUCBA. Available online at https://www.educba.com/careers-in-data-visualization/. Accessed 12/10/2021.

Exhibit 6.30
Sample Executive Summary



Bolt Bank . . . Quick Banking Done Right

Executive Summary

Question

Bolt Bank promises customers quick service. In fact, our commercials promise customers that they will never have to wait more than 40 seconds to see a bank associate. To see if our claims are true, we initiated a business analytics project to answer the question: Are we keeping our promise of a 40-second-or-less wait to customers?

Relevance

The results of this initial business analytics project will allow us to determine if the project needs to be expanded to analyze possible additional hours of operation on weekdays. In addition, the results of this project will help us assess if we need to add additional employees to keep our 40-second promise.

Results

Our analysis was conducted as follows.

(1) Before collecting wait times, we analyzed the number of transactions posted by day and hour. Our analysis revealed that the bank is the busiest from 12–1 p.m. on Fridays. We used this time period to assess customer wait times as it should be associated with the highest wait times given that the number of bank associates working remains constant throughout the workweek.

(2) On Friday, February 7, 2026, we collected 20 random observations of customer wait times from 12 p.m. until 1 p.m.

(3) We analyzed the collected data in several ways.

   a. We first calculated the average wait times for customers during this window as 44.75 seconds, which is higher than our 40-second promise.

   b. We created a histogram in five-second increments to determine the distribution of customer wait times. 11 of the 20 data points, or 55%, were longer than 40 seconds, with the following breakdown:

      i. 40- to 45-second wait: 4 (36%)

      ii. 46- to 50-second wait: 1 (9%)

      iii. 51- to 55-second wait: 3 (27%)

      iv. 56- to 60-second wait: 3 (27%)

(4) The analysis reveals that only 45% of the customers are being served during the busiest hour of the week within the promised 40-second time frame. At least for this time period, Bolt Bank is not keeping its 40-second promise for most of our customers.

Recommendations

Based on our analysis of the busiest hour at Bolt, we recommend the following steps:

(1) Collect customer wait times for an entire week.

(2) Analyze the data using averages and histograms to determine how many customers are waiting longer than 40 seconds.

(3) Based on the additional analysis, determine if additional bank associates need to be hired or if work schedules can be shifted to handle busy periods.

PROGRESS CHECK

23. What is the difference between an executive summary and a full report?
24. What are the components of an executive summary?
25. What are the components of a full report?

Chapter Summary

In this chapter, you learned about the final step in the SOAR model—specifically, how to report the results of your business analytics through data visualization (and sometimes with words). You learned the best types of data visualization to use for numerical versus categorical data and how to use categorical data to provide context and more interesting analyses of your numerical measures. You also learned different methods for visualizing the results of exploratory and confirmatory analytics, as well as how to communicate the results of your analysis with words via executive summaries and reports. No matter which method you use to communicate your results, a good story will help users understand and remember your analytics.

Key Terms

bar chart (*343*) Data visualization that shows categorical data using vertical/horizontal bars with varying heights/lengths proportional to the values they represent.

bins or class intervals (*351*) Subsets of the data, arranged in increasing order, used in a histogram.

categorical data (*342*) Data that are categorized into groups that are represented either by words or by non-meaningful numerical data. For example, categorical data might group people by age (adult, non-adult), identify transaction types (sales versus returns), or specify an inventory costing method (LIFO, FIFO, average cost).

confirmatory analytics (*356*) A type of analytics that uses statistics to judge the likelihood (probability) of a future event or outcome occurring.

continuous data (*347*) Numerical data that can take on any numerical value, not just whole numbers, meaning there is an infinite set of values between any two observations in a continuous data scale. Examples of continuous data are height, weight, and currency.

correlation (*357*) A measure that helps us understand if variables are related to each other and, if so, to what extent.

data visualization (*338*) Means of presenting the results of analysis in an accessible way through charts a graph is a type of chart; sometimes called an infographic.

descriptive statistics (*352*) Summary statistics that briefly describe or summarize features of a data set or variable. The basic descriptive statistics include the mean, standard deviation, minimum, and maximum of a data set.

discrete data (*347*) Numerical data that only take on whole-number (integer) values and have a finite set of values between any two observations. An example of discrete data is the number of products in inventory (you can have 0 items or 150 items in inventory, but you cannot have 147.8 items in inventory).

exploratory analytics (*348*) A type of analytics that uses statistics to summarize and explain performance.

forecast (*358*) A type of analytics that predicts the future.

histogram (*351*) Data visualization that shows the frequency in the distribution of numerical data using bins/intervals arranged in increasing order.

nominal data (*342*) Categorical data that cannot be ranked. Examples include eye color and transaction type (purchase, return).

numerical data (*342*) Meaningful numbers that represent quantities, such as transaction amount, net income, age, or score on an exam.

ordinal data (*342*) Categorical data with natural, ranked categories, such as letter grades (A, B, C, D, F) and Olympic medals (gold, silver, bronze).

pie chart (*343*) Circular graph in which each slice represents the category's proportion of the whole.

proportion (*343*) The number of observations in one category divided by the grand total of all observations.

ratio data (*350*) Numerical data with an equal and definitive value between each data point and a meaningful zero.

regression (*358*) A predictive analytics technique used to estimate an outcome based on variable inputs.

scatterplot (*347*) A graph that plots the values of two variables each on its own axis to identify possible correlation or to identify a trend line or line of best fit.

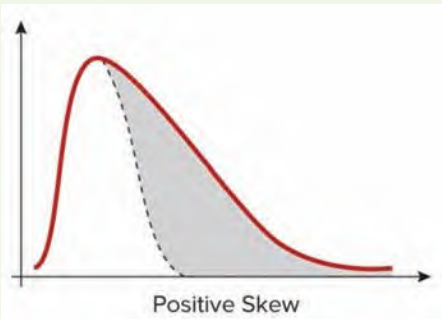
stacked bar chart (*343*) Bar charts with the bars subdivided into different categories.

time series data (349) Values taken on by a variable at different points in time.

trend (358) General direction in which a variable is moving.

ANSWERS TO PROGRESS CHECKS

1. The vertical axis typically shows the values.
2. The horizontal axis typically shows the labels.
3. To sort data in Excel, ensure the active cell is in the column you wish to sort, click the Data tab in the ribbon, and select A-Z (ascending) or Z-A (descending) in the Sort & Filter group.
4. We sorted the data by the numerical value because we were interested in seeing which company had the highest sales, so the numerical ordering is helpful. Sorting by company name would change the order to disregard the numerical values (total sales) and sort the companies in alphabetical order instead (which may not be useful because the emphasis is on sales).
5. A word cloud.
6. Answers will vary. One possible answer: Data visualizations should present the data as they really are. However, some data visualizations are manipulated to support a particular viewpoint. Therefore, we must carefully analyze the visualization to make sure that it presents the facts and only the facts.
7. Line charts are useful when you are working with numerical time series data and when you are more interested in the overall trend of the data instead of specific data points. Line charts also lend themselves to visualizing numerical data that fall below 0.
8. Bar charts and line charts are best for time series data.
9. Pie charts can depict categorical data because they show counts and proportion. They can also be used to depict ranges of numerical data, but bar charts are generally more useful and easier to interpret for this purpose.
10. Histograms represent bins/intervals, while bar charts can represent categorical data. Histograms can only represent numerical data, and their vertical axis shows the count (frequency) of observations. When a bar chart depicts numerical data, it typically shows a value, such as Total Sales, on the vertical axis. A bar chart has space between the bars, but a histogram does not.
11. The vertical axis on a histogram depicts count of observations (frequency) per bin.
12. A positively skewed exam implies the exam is difficult. The bulk of the observations are on the lower end of the distribution, with only a few outliers on the high end. Those high grades are pulling or skewing data to the right, or the positive side of the distribution.



13. 20 students earned a C.
14. 110 students earned above a C.
15. In a box plot, the “X” represents the mean, and the line represents the median.
16. Answers vary. Here is one suggested answer: Skewness, outliers, and median are examples of items that can be identified in a box plot.
17. Creating a box plot will yield more information (and more accurately) than eyeballing a data set because a box plot uses statistics to calculate skewness and outliers.
18. Answers will vary. One possible correct answer: Items that tend to be positively correlated include the amount of time you study and the grades you earn (keep in mind that correlation is not causation), weight and height of newborn babies, and the amount of time spent working out and the amount of weight you lose.
19. Answers will vary. One possible correct answer: Items that tend to be negatively correlated include the number of classes you miss and the grade you earn, and the speed at which you drive and the amount of time it takes to arrive at your final destination.
20. If the correlation between two variables has a value of 0, then the two variables are not correlated (that is, not related at all).
21. An interactive dashboard allows users to filter the data to gain more insights into the data. Often the filter works on multiple worksheets so the user can obtain insights about the data.
22. A single data visualization typically presents one measure of interest. A dashboard presents multiple individual data visualizations at the same time, providing a more holistic view of the company. A dashboard frequently has built-in filters that allow users to see subsets of the data. The filters may affect one or more of the included data visualizations.
23. An executive summary is shorter and simpler than a full report. An executive summary also serves as the first section of a full report.
24. The components of an executive summary are a statement of the problem, a concise description of why the project is relevant to the audience, a summary of the results, and brief recommendations for a course of action that should be taken as a result of the analysis.
25. The components of a full report are an executive summary, a description of the method, an expanded summary of the results (including pertinent data outputs and data visualizations), and expanded recommendations that reiterate the significance of the project.



Multiple-Choice Questions

1. (LO 6.2) Which of the following statements is true?
   a. Line charts should be shown only in descending order to emphasize trend.
   b. Line charts should be shown only in chronological order to emphasize trend.
   c. Line charts should be shown only in alphabetical order to emphasize proportion.
   d. Line charts should be shown only in ascending order to emphasize proportion.

2. (LO 6.2) Pie charts lend themselves well to which type of data showing proportions?
   a. Categorical
   b. Numerical
   c. Diagnostic
   d. Predictive
3. (LO 6.2) Which of the following data are the values taken on by a variable at different points in time?
   a. Cross-sectional data
   b. Categorical data
   c. Numerical data
   d. Times series data
4. (LO 6.4) Which of the following statements is true?
   a. Scatterplots show values of two variables each plotted on its own axis to exhibit potential trends.
   b. Scatterplots show values of two variables each plotted on its own axis to exhibit potential correlation.
   c. Graphs show values of two variables each plotted on its own axis to exhibit potential correlation.
   d. Graphs show values of two variables each plotted on its own axis to exhibit potential trends.
5. (LO 6.3) Exploratory, diagnostic analytics includes which type of visualizations?
   a. Box plots
   b. Scatterplots
   c. Pivot charts
   d. Times series graphs
6. (LO 6.4) Scatterplots are more meaningful if they include which of the following?
   a. Times series prediction
   b. Box plot
   c. Trend line
   d. Categorical variable
7. (LO 6.3) Which of the following is a data point outside the maximum or minimum value of the expected range of a box plot?
   a. Outlier
   b. Median
   c. Mean
   d. Variation
8. (LO 6.2) Accounting financial statement data, such as sales and net income, are most likely to be shown as which type of data?
   a. Interval data
   b. Ratio data
   c. Nominal data
   d. Ordinal data
9. (LO 6.1) What are the individual data observations in a data series called?
   a. Tick marks
   b. Data series
   c. Data points
   d. Markers

10. (LO 6.2) When the trend is the most important item to communicate, which of the following statements is true?
    a. Bar charts are preferable to line charts.
    b. Line charts are preferable to bar charts.
    c. Bar charts are preferable to pie charts.
    d. Pie charts are preferable to line charts.
11. (LO 6.2) Which of the following statements is true?
    a. Although bar charts can represent counts of categorical data, histograms are used only for numerical data.
    b. Although histograms can represent counts of categorical data, bar charts are used only for numerical data.
    c. Although line charts can represent counts of categorical data, bar charts are used only for numerical data.
    d. Although bar charts can represent counts of categorical data, pie charts are used only for numerical data.
12. (LO 6.2) Assume that **Keurig Dr Pepper** launches a new advertising campaign. To understand the impact of the campaign, the company asks a group of customers to provide three words describing the new campaign. Which of the following charts is the best way to identify the words most frequently used to describe the campaign?
    a. Bar chart
    b. Pie chart
    c. Symbol map
    d. Word cloud
13. (LO 6.2) Assume that **Keurig Dr Pepper** wants to create a data visualization that compares the total number of cans of Dr Pepper sold in each U.S. state. Which of the following charts is BEST way to present the relative differences in cans sold?
    a. Bar chart
    b. Filled geographic map
    c. Heat map
    d. Tree map
14. (LO 6.3) The stock market typically yields a small positive return each day, but intermittently it yields a high negative return on a few days. What type of distribution describes this behavior?
    a. Negatively skewed
    b. Symmetric
    c. Positively skewed
    d. Uniform
15. (LO 6.3) A company is considering a stock investment that is (projected to be) positively skewed. Which of the following statements is correct about the investment?
    a. The company should expect frequent small losses and a few large gains from the investment.
    b. The company should expect frequent large losses and a few small gains from the investment.
    c. The company should expect losses and gains equally.
    d. The company should expect few small losses and frequent large gains from the investment.
16. (LO 6.5) During weekly executive meetings, a CEO likes to see how visualizations change if different filters are applied to the data. Business analysts who prepare the

dashboard do not know ahead of time which filters are needed. Which type of dashboard should business analysts create, so that the CEO can change the filters as many times as desired during the meetings?

a. Cross-section
b. Interactive
c. Static
d. Trend

17. (LO 6.6) What does the target audience for business analytics help to determine?
    a. Types of charts used
    b. Language and terms used
    c. Type of report used
    d. Types of charts and reports used as well as the language and terms used

Discussion Questions

1. (LO 6.2) Why is a bar chart optimal for visualizing numerical data and a pie chart optimal for visualizing categorical data?
2. (LO 6.1, 6.2) How does sorting help decision-makers more quickly grasp the information conveyed in a bar chart?
3. (LO 6.1, 6.2, 6.5) When should you use a table instead of a bar chart?
4. (LO 6.3) In what way is a box plot more informative than a histogram?
5. (LO 6.4) Why is a line chart a good option for visualizing a forecast?
6. (LO 6.6) What are the four components of an executive summary? Why is it important to keep the executive summary to one page or less?
7. (LO 6.5) What is a dashboard? How does it help decision-makers?
8. (LO 6.1, 6.2) Why are data visualizations a good way to present large amounts of data?



Brief Exercises

1. (LO 6.1, 6.2, 6.3, 6.4, 6.5) For each example, which descriptive data visualization is better: a pie chart or a bar chart or line chart?

Example	Data Visualization (Pie Chart or Bar/Line Chart)
Trend of net income over the past 5 years	
Proportions (relative sales of each of the company's products to the company's total sales)	
Ratio data (data with a meaningful zero)	
Company sales versus returns	
Categorical data (number of companies using different depreciation methods)	

2. (LO 6.1, 6.2, 6.3, 6.4, 6.5) Match each term with its definition.

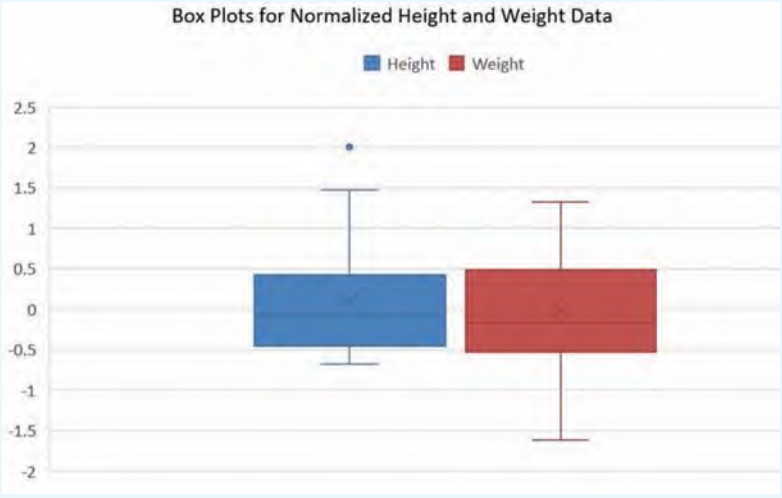
Term	Definition
1. Descriptive statistics	a. The values of two variables, each plotted on its own axis to exhibit possible correlation
2. Trend	b. The number of observations in a particular category divided by the grand total of observations
3. Proportion	c. General direction in which something is moving
4. Ratio data	d. Numerical data with an equal and definitive value between each data point and a meaningful zero
5. Scatterplot	e. Values taken on by a variable at different points in time
6. Time series data	f. Brief summaries of a data set that reflect the data set as a whole

3. (LO 6.1, 6.2, 6.3, 6.4, 6.5) Match the data visualization with the type of business analytics (descriptive, diagnostic, predictive, or prescriptive) it is most associated with.

Data Visualization	Type of Business Analytics (Descriptive, Diagnostic, Predictive, or Prescriptive)
Scatterplot	
Times series data	
Tables displaying sensitivity analysis	
Pie charts	
Line and bar charts	
Box plots	
Histograms	

Problems

1. (LO 6.3) The following box-and-whisker plots illustrate the heights (blue) and weights (red) of a sample of children. To make it easier to compare the two box plots, the heights and weights of each individual have been normalized or transformed to a *z*-score, which sets different data sets on the same scale. For a refresher on *z*-scores, please see Chapter 3.

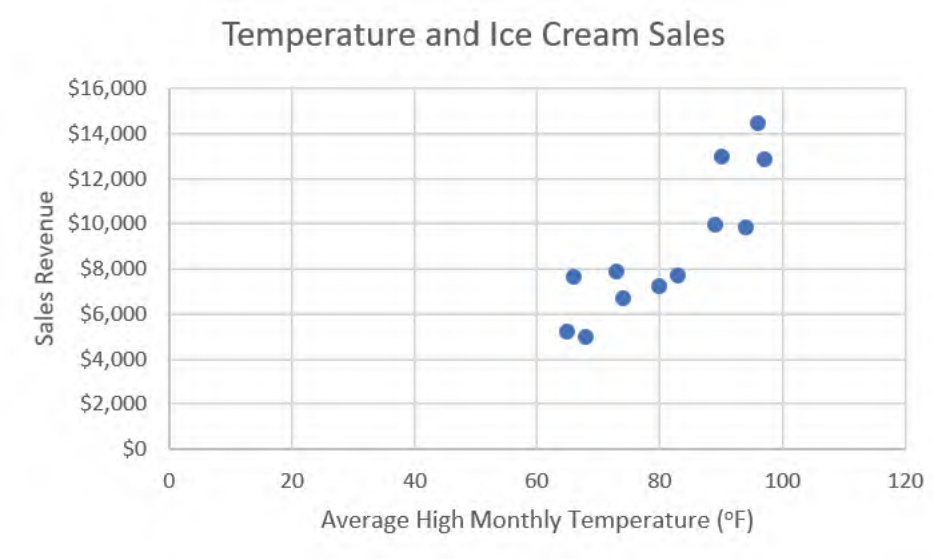


a. Does either of the box plots have an outlier? If so, which one? Is the outlier positive or negative?

b. If there is an outlier and we assume that it is not an error in the data set, then what does the outlier imply about the individual in comparison to the rest of the individuals sampled?

c. Which set of data, height or weight, has wider variability (excluding any outliers)?

2. (LO 6.4) The following scatterplot shows the relationship between average high monthly temperature (in Fahrenheit) and ice cream sales for a small ice cream stand. The correlation coefficient is 0.87.

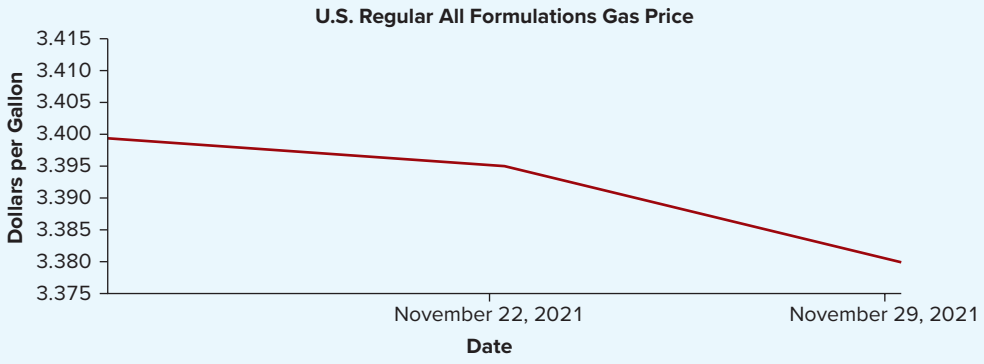


a. Interpret the correlation value. Is the correlation strong?

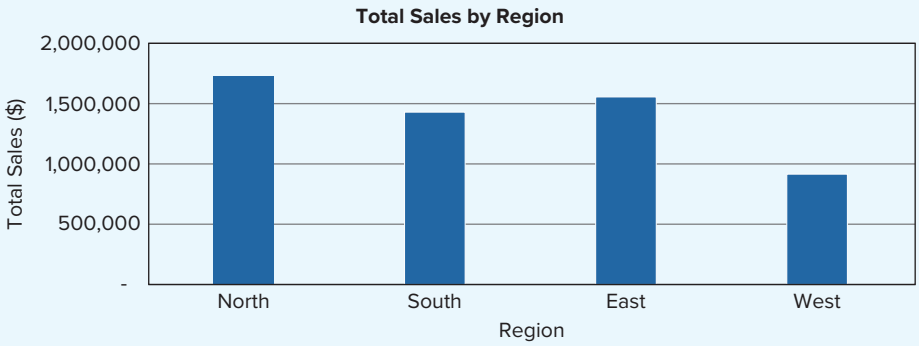
b. Using the scatterplot, determine if the two variables are positively correlated or negatively correlated.

c. Suppose we want to run a linear regression model/equation for the data. What is the independent variable? What is the dependent (outcome) variable?

3. (LO 6.5, 6.6) On December 2, 2021, the following graph appeared in a tweet. The graph shows the decline of regular unleaded gas prices in the United States between November 22 and November 29, 2021. Analyze the data visualization with respect to changes in the price of gas. What suggestions, if any, do you have for improving the graph?



4. (LO 6.3) Examine the data visualization below. Is it a histogram or a bar chart? How can you tell?



Use the following information to answer Problems 5–10.
HackUsNow (a fictional company) conducts scientific research on alternative fuels. Like all modern companies, its computers are connected to the Internet. The company's employees work in one of three departments: Scientific, Administrator, or Secretary. The Scientific department conducts the research. The Administrator department oversees the company's operations. The Secretary department prepares research grants and research reports. The company currently has three office locations: Connecticut, New Jersey, and New York.

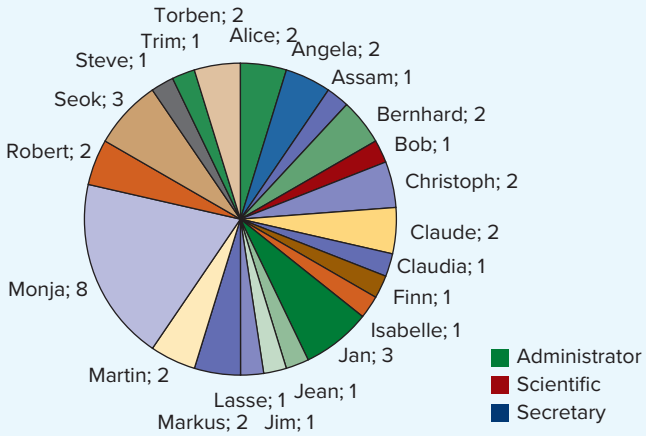
HackUsNow is concerned about its computers being hacked because of the proprietary nature of its research (and the potential for another company to profit from it). For this reason, the company wants to investigate its computer hacks.

Data for HackUsNow are contained in the data file HackUsNow.xlsx. Problem 5 asks a question about a prepared data visualization for HackUsNow. Use the HackUsNow.xlsx file to answer Problems 6 through 10. The data dictionary for the file follows.

Data Dictionary

Employee:	Employee first name
Hacks:	The number of times the employee's computer has been hacked
Department:	The department in which the employee works
Computer Age:	New or Old (New is less than three years old)
Location:	Location of office (Connecticut, New Jersey, New York)

5. (LO 6.2, 6.3) The following pie chart shows how frequently each employee's computer at HackUsNow has been hacked. Evaluate the pie chart assuming that the analyst is trying to answer this question: "Which employees' computers were hacked the most?" Does the pie chart answer the question? What recommendations do you suggest for improving this pie chart?



6. (LO 6.1, 6.2, 6.3) Problem 5 presents a pie chart for HackUsNow. Using Tableau, create a bar chart sorted by the number of hacks using the data in HackUsNow.xlsx. Does the bar chart clearly answer the question "Which employees' computers were hacked the most?"?
7. (LO 6.2) Problem 5 presents a pie chart for HackUsNow. Using Tableau, create a tree map using the data in HackUsNow.xlsx. The tree map should show the employees' names and number of hacks to clearly answer the question "Which employees' computers were hacked the most?"
8. (LO 6.2) Using the data in HackUsNow.xlsx, create a data visualization in Tableau to clearly answer the question "Which employees experienced more than the average number of hacks?"
9. (LO 6.2) Using the HackUsNow.xlsx data, create a data visualization in Tableau to answer the question "Which location has the most hacks?"
10. (LO 6.2, 6.3) The data set HackUsNow.xlsx includes an additional column labeled Computer Age that has not been analyzed yet. Ask a question related to Computer Age and hacks. Then create a data visualization in Tableau to answer that question. What conclusion, if any, did you make based on the analysis?

Data Visualization Exercise

1. Refer to Problems 5–10 above. Assume that the question that HackUsNow wants to answer is "Which department's computers have been hacked the most frequently?" That is, the company wants to know total hacks by department. We can't easily determine the answer to that question based on the pie chart presented in Problem 5 because too many colors were used. So, let's create a bar chart by department to answer the question. Fortunately, department data are contained in the HackUsNow.xlsx data set.

Now that we have the data, we can start analyzing them. Proceed as follows.

1. Open Tableau.
2. Click Microsoft Excel.
3. Find HackUsNow.xlsx.
4. Click Open.
5. Verify that Location was imported as geographic (a globe should appear above the Location column in the Data Source tab). If not, click on Abc, and select **Geographic Role > State/Province.** Although we are not using Location in this visualization, we may choose to use it in future data visualizations.
6. Click Sheet 1.
7. Drag Department to the Columns shelf.

8. Drag Hacks to the Rows shelf.
9. Drag Department to Colors in the Marks card. You should now have three colored columns, one for each department, as shown here:

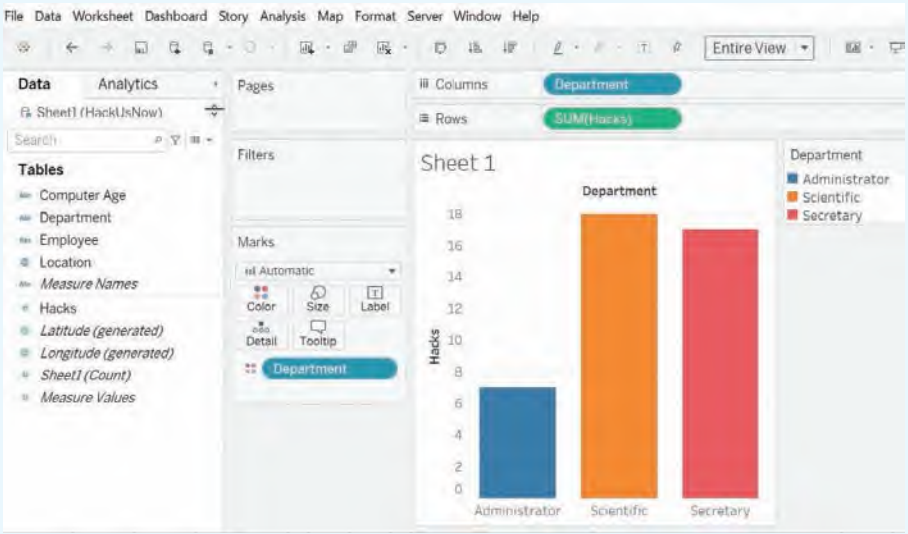


Tableau Software

10. Let's format the chart to make it easier to read. First, sort the columns in descending order. Click on the Sort in descending order icon .
11. Next, expand the graph to fill the entire canvas. Change the Standard dropdown to Fit Width:



Tableau Software

12. Now make the departmental colors match the original colors used in the Problem 5 pie chart legend, where Administrator is green, Scientific is red, and Secretary is blue. To do that, click on the down arrow in the right corner of the legend", which is in a separate box at the right corner next to the canvas/view with the bar chart. Select Edit Colors . . .:

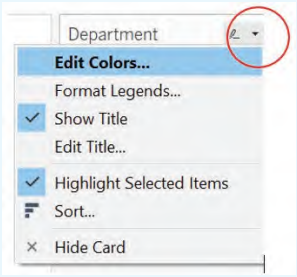


Tableau Software

13. In the Edit Colors . . . [Department] dialog box, select a data item and then select the closest Automatic color. Click OK when colors are correct, as shown here:

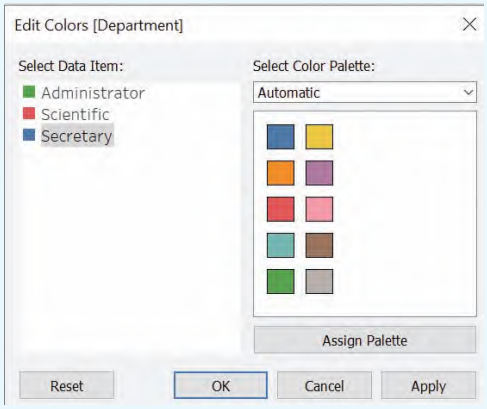


Tableau Software

14. Now rename the visualization. Right-click on the Sheet 1 tab and select Rename. Type in "Hacks by Department" and then press Enter.

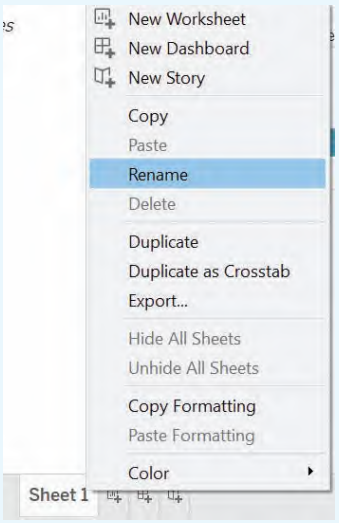


Tableau Software

15. We have created a visualization that clearly answers HackUsNow's question: "Which department's computers have been hacked the most frequently?" The answer, as shown in the following bar graph, is the Scientific department.



Tableau Software

LAB 6.1 **Excel:** Descriptive Analytics: Visualizing Pivot Table Data Using Conditional Formatting and Sparklines

LAB 6.2 **Tableau:** Descriptive Analytics: Analyzing Sales Revenue by with a Tree Map

LAB 6.2 **Power BI:** Descriptive Analytics: Analyzing Sales Revenue by with a Tree Map

LAB 6.3 **Tableau:** Descriptive Analytics: Analyzing Sales Revenue by with a Bar Chart and Filters

LAB 6.3 **Power BI:** Descriptive Analytics: Analyzing Sales Revenue by with a Bar Chart and Filters

LAB 6.4 **Excel:** Descriptive Analytics: Creating a Dashboard Using Pivot Tables and Slicers

LAB 6.5 **Tableau/Power BI:** Descriptive Analytics: Creating a Dashboard

Lab 6.1 Excel

Descriptive Analytics: Visualizing Pivot Table Data Using Conditional Formatting and Sparklines

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Descriptive Analytics, Pivot Table, Sparklines, Conditional Formatting

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" One of the techniques used in descriptive analytics is a pivot table, which is a flexible way to summarize large amounts of data.

Sláinte is a fictional brewery that sells six different beers. It has recently expanded the distribution of its products from one state to nine states, and now its business has begun stabilizing after the expansion. With that stability comes a need for better analysis.

Assume Sláinte has hired you to help management better understand the company's sales data and provide input for its strategic decisions. In this lab, you will analyze sales by product. To do so, you will create a pivot table with conditional formatting and sparklines. Conditional formatting and sparklines are quick ways to visualize and compare data and trends when a full-fledged chart isn't necessary. You will also create a stacked bar chart to help present the results.

Required

1. Create a pivot table.
2. Add conditional formatting for the grand totals.
3. Add sparklines for each type of beer.
4. Create a stacked bar chart from the pivot table results.

Specify the Question

How many cases of each beer are sold each month and in total from January through June?

Obtain the Data

This lab uses the data in Lab 6.1 Data.xlsx, which contains sales transaction data for Sláinte from January through June of 2026.[8] The data dictionary for this file is below. The data include three worksheets/tables: Sales_Transactions, Customers, and Products. In the Sales_Transactions worksheet/table, the data include an observation/row for each sales transaction and product sold. So, if a transaction included the sale of two products, the data set has one unique Sales_Order_ID with two rows (that is, one for each product).

Data Dictionary

Sales_Transactions Table:

Sales_Order_ID:	Unique sales transaction identifier
Sales_Order_Date:	Date of sales order
Sales_Employee_ID:	Unique identification number of the employee making the sale
Customer_ID:	Unique customer identifier
Product_Code:	Unique product identifier
Sales_Order_Quantity_Sold:	Number of cases sold in the sales transaction

Customers Table:

Customer_ID:	Unique customer identifier
Business Name:	Customer name
Customer_Address:	Customer street address
Customer_City:	Customer city
Customer_St:	Customer state
Customer_Zip:	Customer zip code

Products Table:

Product_Code:	Unique product identifier
Product_Description:	Type of beer
Product_Sale_Price:	Price of one case of beer

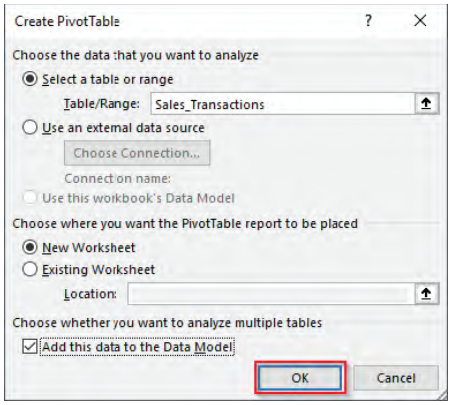
Analyze the Data

Analysis Task 1: Create a Pivot Table.

1. Insert a pivot table from any of the tables in the Excel workbook. To do so, place the cursor inside the data on any worksheet. From the menu, select **Insert > Tables > Pivot Table.** Check the box next to "Add this data to the Data Model," as shown in Lab Exhibit 6.1.1.

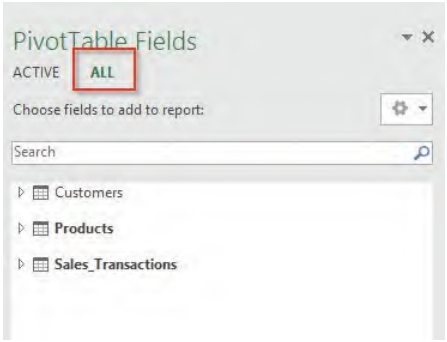
[8]Source: These data are a slightly altered version of Lab 3-4 data in Richardson, Terrell, and Teeter, 2021. *Introduction to Data Analytics for Accounting.* New York, NY: McGraw Hill.

Lab Exhibit 6.1.1



2. In the PivotTable Fields list, click ALL, as shown in Lab Exhibit 6.1.2 to show the fields from each of the three tables in the workbook and click on the arrows to see the fields. The relationships between the tables have already been created in this data set.

Lab Exhibit 6.1.2



3. Place the following fields in your pivot table (as shown in Lab Exhibit 6.1.3):
   - Sales_Order_Quantity_Sold (from the Sales_Transactions sheet) in Values
   - Product_Description (from the Products sheet) in Rows
   - Sales_Order_Date (Month) in Columns

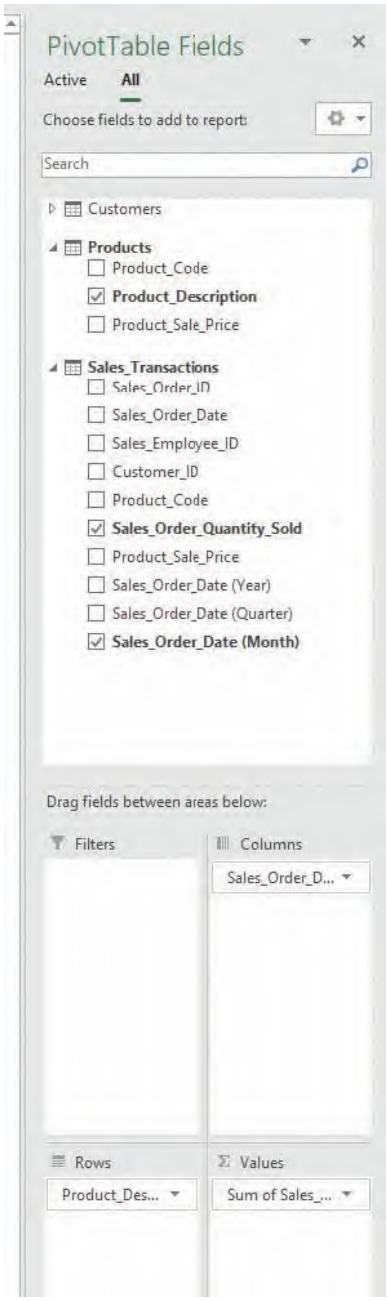
   If prompted to auto-detect relationships, click auto-detect in the PivotTable Fields list window.

Analysis Task 2: Add Conditional Formatting.

To quickly visualize how each product's total quantity sold over time compares across all six of Sláinte's products, we can apply conditional formatting to the Grand Total column.

1. Select the data in the Grand Total column of your pivot table, as shown in Lab Exhibit 6.1.4.

Lab Exhibit 6.1.3



Lab Exhibit 6.1.4

Sum of Sales_Order_Quantity_Sold	Column Labels						
Row Labels	Jan	Feb	Mar	Apr	May	Jun	Grand Total
Imperial IPA	23		2	17	8	11	61
Imperial Stout	61	27	42	36	51	41	258
IPA	36	35	66	34	23	41	235
Pale Ale	27	4	32	18	30	5	116
Stout	65	33	4	16	35	37	190
Wheat	21	11	54		17	16	119
Grand Total	**233**	**110**	**200**	**121**	**164**	**151**	**979**

2. From the menu, select **Home > Styles > Conditional Formatting**. A menu with the different types of formatting available will appear.
3. Select Data Bars and select the first option for blue gradient fill bars. This conditional formatting is helpful because it allows us to compare grand totals of each product.

Analysis Task 3: Add Sparklines.

1. To see how each product's month-over-month sales compare to one another, we can display mini line charts next to each row with a sparkline. To do so, select all of the "meat" of your pivot table—that is, don't select any of the product labels (such as Imperial IPA), month labels, or grand totals (Lab Exhibit 6.1.5).

Lab Exhibit 6.1.5

Sum of Sales_Order_Quantity_Sold	Column Labels						
Row Labels	Jan	Feb	Mar	Apr	May	Jun	Grand Total
Imperial IPA	23		2	17	8	11	61
Imperial Stout	61	27	42	36	51	41	258
IPA	36	35	66	34	23	41	235
Pale Ale	27	4	32	18	30	5	116
Stout	65	33	4	16	35	37	190
Wheat	21	11	54		17	16	119
Grand Total	**233**	**110**	**200**	**121**	**164**	**151**	**979**

2. From the menu, select **Insert > Sparklines > Line** to add sparklines to each product.
3. A window will appear, specifying the data range you just selected and awaiting input for the Location Range. We'd like to see the trend lines to the immediate right of our PivotTable, so select the cells in the first empty column after your Grand Totals (as shown in Lab Exhibit 6.1.6).

Lab Exhibit 6.1.6

3	Sum of Sales_Order_Quantity_Sold	Column Labels						
4	Row Labels	Jan	Feb	Mar	Apr	May	Jun	Grand Total
5	Imperial IPA	23		2	17	8	11	61
6	Imperial Stout	61	27	42	36	51	41	258
7	IPA	36	35	66	34	23	41	235
8	Pale Ale	27	4	32	18	30	5	116
9	Stout	65	33	4	16	35	37	190
10	Wheat	21	11	54		17	16	119
11	**Grand Total**	**233**	**110**	**200**	**121**	**164**	**151**	**979**

4. Click OK, and your sparklines will be created.
5. You will notice that there are gaps in the lines, though. If you'd rather see a continuous line to represent zero values for the blank cells, you can change this option. Ensure that one of the cells with the sparkline in it is active, and navigate to the Sparkline Tools tab above the menu.
6. Click the down arrow in the Edit Data button, and then select Hidden & Empty Cells.
7. In the Hidden and Empty Cell Settings window, select the option to show empty cells as zero, and click OK.

Report the Results: Create a Stacked Bar Chart



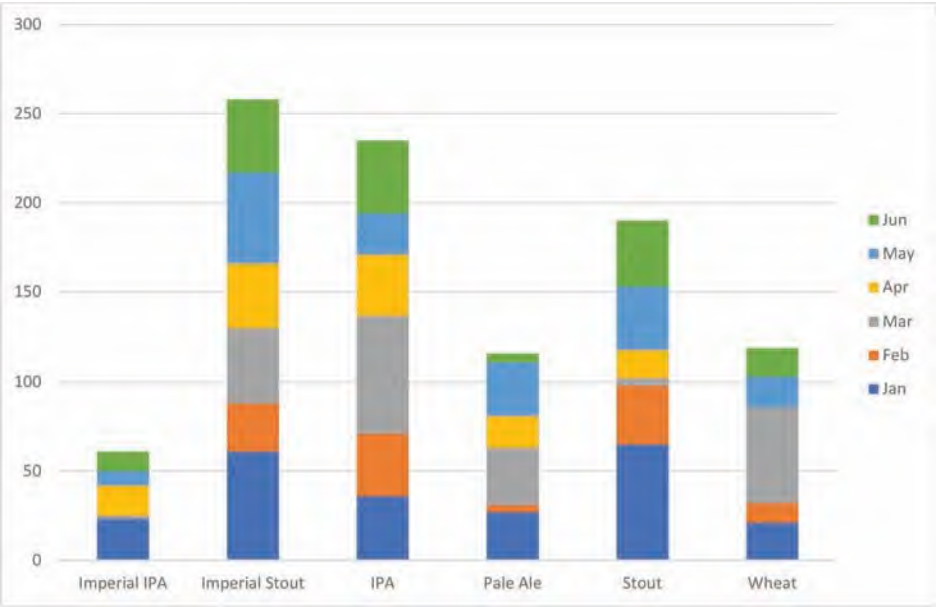
The combination of the conditional formatting and the sparklines helps to tell the story of how sales have performed across products during the first six months of 2026. The conditional formatting in the Grand Total column provides a quick, easy summary of how each product performed overall. The sparklines offer a quick, easy summary of how each beer

performed month over month. As you prepare for your presentation to management, you decide to make a stacked bar chart. To do so, proceed as follows:

1. Place the cursor in the pivot table.
2. From the PivotTable Tools menu, click **Analyze > Tools > PivotChart.** Then select **Column > Stacked Column** (the second graph).
3. Click OK.

An interactive stacked bar chart (Lab Exhibit 6.1.7) is created that you can use to answer questions when you present results to management.

Lab Exhibit 6.1.7



Assessment

1. Take a screenshot of your pivot table (with the conditional formatting and sparklines), paste it into a Word document named "Lab 6.1 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the completed stacked bar chart, paste it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Lab 6.1 Excel Submission.docx via Connect, if assigned.

Alternate Lab 6.1 Excel

Descriptive Analytics: Visualizing Pivot Table Data Using Conditional Formatting and Sparklines

Apply the same steps in Lab 6.1 to the Alt Lab 6.1 Data.xlsx data set. While Lab 6.1 uses data for transactions with customers, the Alt Lab 6.1 data set contains data regarding purchasing transactions with vendors/suppliers for raw materials, which will be converted into the beer that Sláinte will sell to customers. A new data dictionary is presented in the "Obtain the Data" section.

Specify the Question

What quantity of raw materials is purchased each month?

Obtain the Data

This lab uses the data in Alt Lab 6.1 Data.xlsx, which contains supplier/vendor transaction data for Sláinte from January through June of 2026.[9] The data dictionary for this file appears below. The data include three worksheets/tables: Purchase_Orders, Suppliers, and Raw_Materials. In the Purchase_Orders worksheet/table, the data include an observation/row for each purchase transaction and type of raw material purchased. So, if a transaction included the purchase of two different raw materials, the data set will have one unique Purchase_Order_ID with two rows (that is, one for each raw material type).

Data Dictionary

Purchase_Orders Table:

Purchase_Order_ID:	Unique purchase transaction identifier
Purchase_Order_Date:	Date of purchase order
Purchase_Employee_ID:	Unique identification number of the employee making the purchase
Supplier_ID:	Unique supplier/vendor identifier
Raw_Materials_Code:	Unique raw material product identifier
Quantity_Purchased:	Quantity of raw material purchased in the transaction

Suppliers Table:

Supplier_ID:	Unique supplier/vendor identifier
Supplier_Company_Name:	Supplier/vendor name
Supplier_Address:	Supplier/vendor street address
Supplier_City:	Supplier/vendor city
Supplier_St:	Supplier/vendor state
Supplier_Zip:	Supplier/vendor zip code

Raw_Materials Table:

Raw_Materials_Code:	Unique raw material product identifier
Raw_Materials_Description:	Description of raw material
Raw_Materials_Price:	Price of one unit of raw materials

Required

1. Create a pivot table to show the quantity of each type of raw material purchased per month.
2. Add conditional formatting for the grand totals.
3. Add sparklines for each product. If necessary, adjust your sparklines so that you can see a continuous line to represent zero values.
4. Create a stacked bar chart for the pivot table results.

Assessment

1. Take a screenshot of your pivot table (with the conditional formatting and sparklines), copy it into a Word document named "Alt Lab 6.1 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the completed stacked bar chart, paste it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Alt Lab 6.1 Excel Submission.docx via Connect, if assigned.

[9]These data are a slightly altered version of Lab 4-5 data in Richardson, Terrell, and Teeter, 2021, *Introduction to Data Analytics for Accounting.* New York, NY: McGraw Hill.

Lab 6.2 Tableau

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Descriptive Analytics: Analyzing Sales Revenue by Product with a Tree Map

Keywords

Tree Map, Calculated Field, TODAY() Function

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" As discussed in the text, the type of data determines the best type of visualizations. In this lab, we analyze proportional data, with proportion defined as the number of observations in one category divided by the grand total of all observations. Using proportions is one of the primary methods of summarizing categorical data. One way to visualize proportional data is a tree map, which uses size and color to show the proportional size of values.

Sláinte is a fictional brewery that sells six different beers. It has recently expanded the distribution of its products from one state to nine states, and now its business has begun stabilizing after the expansion. With that stability comes a need for better analysis.

Assume Sláinte has hired you to help management better understand the company's sales data and provide input for its strategic decisions. Management wants to know total sales revenue for the period as well as the proportion of total sales revenue for each beer. In this lab, you will create a tree map for Sláinte using Tableau to break down total sales revenue per product sold by Sláinte. The type of beer is categorical data.

When you import data into Tableau, it automatically assigns a data type (such as name, date, or number) to each field. As part of the ETL (extract, load, and transform) process, these data types need to be verified and changed if necessary. Based on the data type, Tableau assigns a role to the field: either a dimension or a measure.

- *Measures* are numeric fields that can be used for mathematical calculations.
- *Dimensions* are non-numeric and used to categorize data. Dimensions include fields identified as string (text), dates, and geographic data types.

To create a tree map in Tableau, the visualization needs at least one measure and one dimension.

Required

1. Calculate total sales revenue.
2. Create a tree map for sales revenue broken down by beer, showing the proportion of sales revenue attributable to each beer.
3. Create a dashboard showing the tree map, total sales revenue, and date of the report (today's date). The dashboard is static, or not interactive, meaning that it does not have built-in filters to change the display. This dashboard creates a report to present the results of analytics.

Specify the Questions

1. What is total sales revenue?
2. What proportion of total sales revenue did each beer generate?

Obtain the Data

This lab uses the data in Lab 6.2 Data.xlsx, which contains sales transaction data for Sláinte from January through June of 2026.[10] The data dictionary for this file appears below. The

data include three worksheets/tables: Sales_Transactions, Customers, and Products. In the Sales_Transactions worksheet/table, the data include an observation/row for each sales transaction and product sold. So, if a transaction included the sale of two products, the data set will have one unique Sales_Order_ID with two rows (that is, one for each product).

Data Dictionary

Sales_Transactions Table:

Sales_Order_ID:	Unique sales transaction identifier
Sales_Order_Date:	Date of sales order
Sales_Employee_ID:	Unique identification number of the employee making the sale
Customer_ID:	Unique customer identifier
Product_Code:	Unique product identifier
Sales_Order_Quantity_Sold:	Number of cases sold in the sales transaction

Customers Table:

Customer_ID:	Unique customer identifier
Business Name:	Customer name
Customer_Address:	Customer street address
Customer_City:	Customer city
Customer_St:	Customer state
Customer_Zip:	Customer zip code

Products Table:

Product_Code:	Unique product identifier
Product_Description:	Type of beer
Product_Sale_Price:	Price of one case of beer

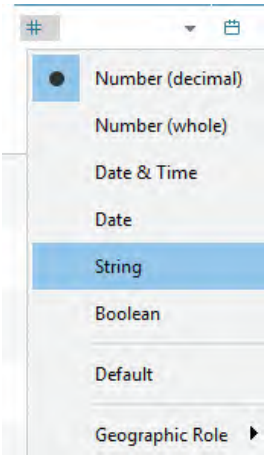
Obtain Task 1: Open Tableau and Prepare the Data.

1. Open Tableau.
2. Under Connect, select Microsoft Excel.
3. In the Open window, browse to the Lab 6.2 Data.xlsx file.
4. Click Open.
5. In Tableau 2020.2 and higher, the data source pane defaults to a logical layer for building relationships. For this lab, we will use Tableau's default relationships. Drag Sales_Transactions to the Data Connection page.
6. Drag Products to the Data Connection page. In the Edit Relationship dialog box, verify that [Product Code] = [Product Code (Products)]. Click x to close the dialog box.
7. Click on the Sales_Transaction rectangle. If Tableau imported Sales Order ID as a number, change the data type to string because it is a text identifier, not a numeric number for calculations. To change the type, click on the # above Sales Order ID and change it to string, as shown in Lab Exhibit 6.2T.1.

[10]These data are a slightly altered version of Lab 3-4 data in Richardson, Terrell, and Teeter, 2021, *Introduction to Data Analytics for Accounting.* New York, NY: McGraw-Hill.

Lab Exhibit 6.2T.1

Tableau Software



8. Change Sales Employee ID, Customer ID and Product Code to a string data type.
9. Click on the Products rectangle. Verify that Product Code (Products) is a string data type. Change if necessary. The data are ready to use.

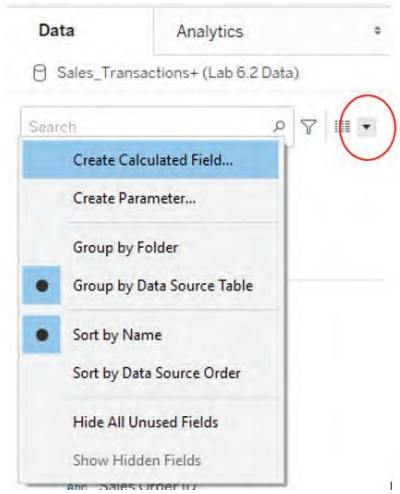
Analyze the Data

Analysis Task 1: Calculate Total Sales Revenue.

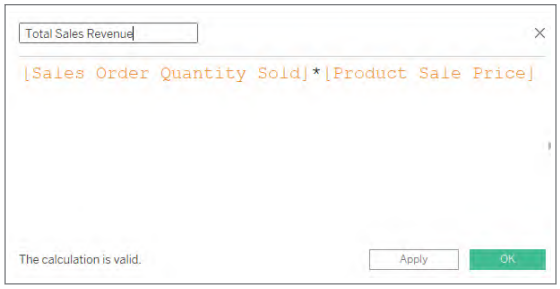
1. Click on Sheet 1.
2. To determine total sales revenue, create a calculated field as follows: Click on the down arrow at the top of the Data tab. Select Create Calculated Field. . ., as shown in Lab Exhibit 6.2T.2.

Lab Exhibit 6.2T.2

Tableau Software



3. In the dialog box that appears, type "Total Sales Revenue" (over the grayed Calculation1) as the variable name. Type the formula for total sales, which equals number of units sold times sales price, or Sales Order Quantity Sold * Product Sale Price. As you type in the variable names, Tableau shows the matching variables; select the appropriate variable name. Lab Exhibit 6.2T.3 shows the dialog box with the formula.



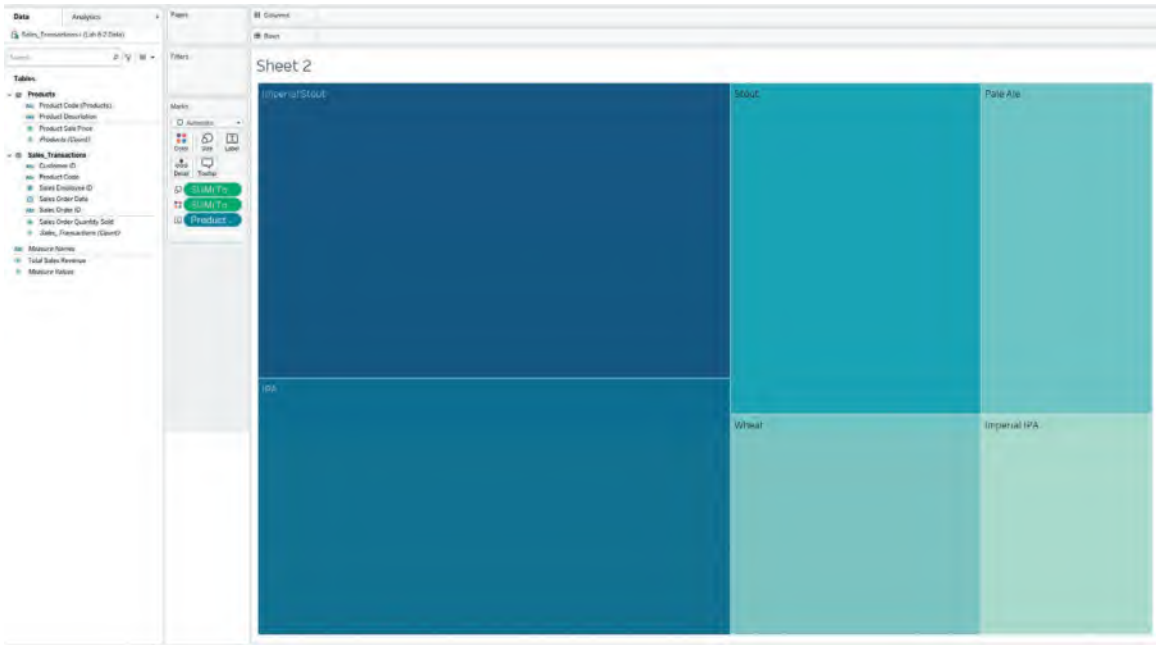
Lab Exhibit 6.2T.3

Tableau Software

4. Verify that the bottom-left corner of the box shows "The calculation is valid." If so, click OK. If not, correct any errors. The new variable, Total Sales Revenue, should appear in the Data tab towards the bottom with the Measures.
5. Drag Total Sales Revenue to the canvas. The total sales revenue amount of 95,790 should appear.
6. To format total sales revenue as currency, click the down arrow in the green pill in the Marks card. Select Format. In the Format Pane at the left, change Default Numbers to Currency (Standard), English (United States).
7. Rename the worksheet to Total. Double-click on the Sheet 1 tab, type "Total", and press enter.

Analysis Task 2: Create a Tree Map of Sales Revenue by Beer.

1. Add a new sheet. Click the new Worksheet button at the bottom of the screen.
2. Double-click on Product Description.
3. Double-click on Total Sales Revenue.
4. If a tree map does not appear automatically, select tree map on the Show Me pane located on the top right of the screen. Your sheet should look like Lab Exhibit 6.2T.4.



Lab Exhibit 6.2T.4

Tableau Software

5. To see the total sales revenue by beer, drag Total Sales Revenue to Label in the Marks card.
6. To calculate the proportion/percentage of sales revenue by beer, complete the following steps: (a) drag Total Sales Revenue below the existing SUM(Total Sales Revenue) in the Marks card that has a label T marker in front of it; (b) click the down arrow on the new pill; (c) select Quick Table Calculations; and (d) select Percent of Total. Lab Exhibit 6.2T.5 shows the steps.

Lab Exhibit 6.2T.5

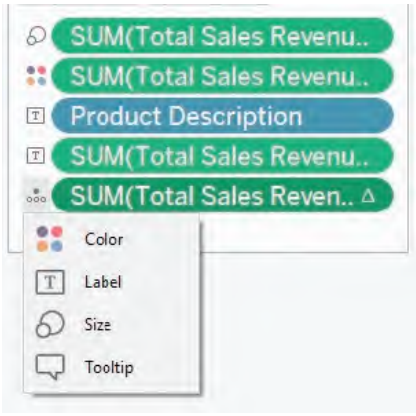
Tableau Software



7. To format sales revenue as currency, click the down arrow in the SUM(Total Sales Revenue) green pill in the Marks card with a Label. Select Format. In the Format Pane at the left, change Default Numbers to Currency (Standard), English (United States).
8. To see the proportion in the data visualization, you need to change the proportion calculation pill from detail (. . .) to label. Click on detail, and from the drop-down menu select Label, as shown in Lab Exhibit 6.2T.6.

Lab Exhibit 6.2T.6

Tableau Software



Analysis Task 3: Label the Sheet with Today's Date as the Report Date.

When analyzing data, it is important to know what date the report was run. Let's use today's date as the report date.

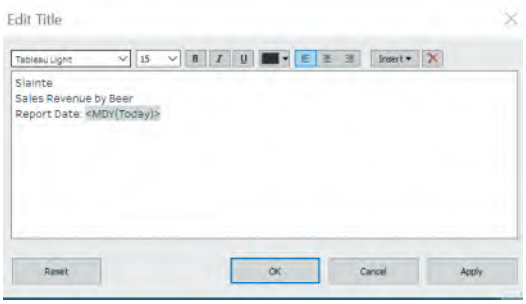
1. Create a new calculated field called Today with the formula equal to TODAY().
2. Drag Today onto Detail in the Marks card. Right-click YEAR(Today). Select **More > Custom.** In the Custom Date box, use the drop-down arrow to select Month/Day/Year, as shown in Lab Exhibit 6.2T.7. Click OK.



Lab Exhibit 6.2T.7

Tableau Software

3. To create a sheet title, right-click on Sheet 2 title at the top-left corner of the view (where the tree map is located). Select Edit Title.
4. In the Edit Title dialog box, on the first line type "Slainte". On the second line, type "Sales Revenue by Beer". On the third line, type "Report Date:". Click on the down arrow by Insert. Select MDY(Today). The Edit Title Box should look like Lab Exhibit 6.2T.8. Click OK.



Lab Exhibit 6.2T.8

Tableau Software

The sheet should now have a title with the report date, which will change with the current date.

5. The legend is labeled SUM(Total Sales Revenue) and has a range of 7,320 to 27,090. Remove the legend on the top right because it does not add any additional information. Click the right down arrow in the legend by SUM(Total Sales Revenue). Select Hide Card.
6. Rename the worksheet. Double-click on the Sheet 2 tab, type "Sales Revenue by Beer", and press Enter.

Report the Results: Create a Dashboard for the Results

You just created two sheets in Tableau. The first sheet shows total sales revenue. The second sheet shows the revenue breakdown by beer using a tree map with total sales revenue and proportion of sales for each beer. To create a professionally formatted report showing all of the data in one place, you can use a dashboard to combine the two worksheets. Dashboards help you share a story, providing the information in an easy-to-use and professional format for decision-making.

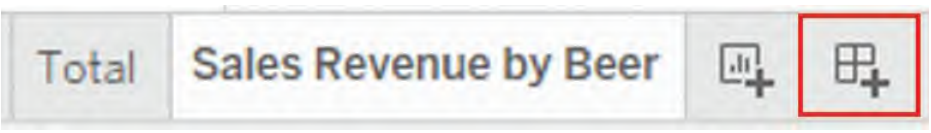


DATA VISUALIZATION

1. Click on the new dashboard button, which is outlined in red in Lab Exhibit 6.2T.9.

Lab Exhibit 6.2T.9

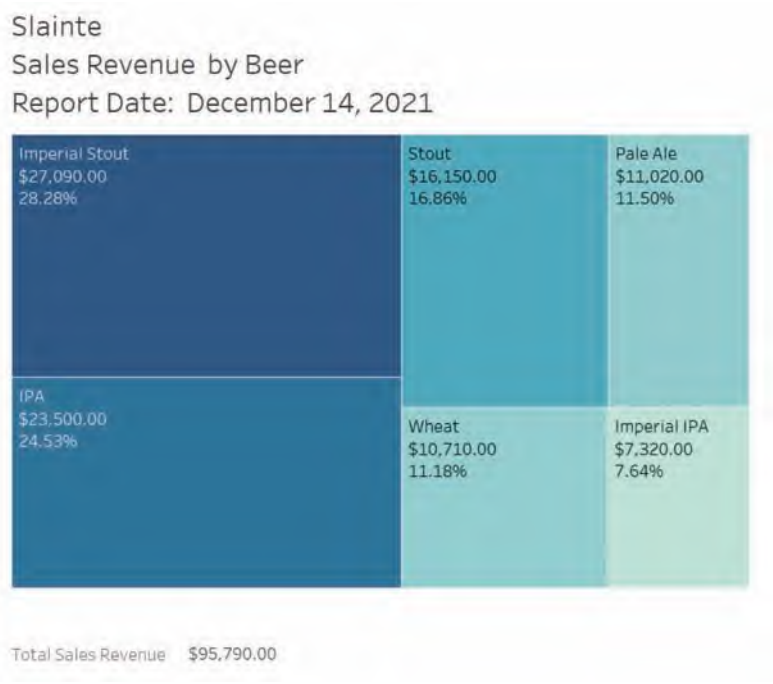
Tableau Software



2. Make sure Tiled is clicked at the bottom left of the dashboard tab.
3. Drag the Sales Revenue by Beer sheet to where it says Drop Sheets here.
4. Click on Floating at the bottom left of the dashboard tab. Floating gives more options for arranging the sheets on the dashboard.
5. Add a label for total sales revenue. Under Objects, double-click Text. In the Edit Text dialog box, type "Total Sales Revenue". Click OK.
6. Drag the Total Sales Revenue box below the tree map.
7. Drag the Total sheet to the right of the Total Sales Revenue box. Right-click on Total. Select Hide Title. . .
8. Resize the Total Sales Revenue box and place the total next to the Text box.
9. Name your dashboard "Sales Report". Your dashboard should look like Lab Exhibit 6.2T.10. *Note:* The Report Date will be the current date.

Lab Exhibit 6.2T.10

Tableau Software



Assessment

1. Take a screenshot of your dashboard, copy it into a Word document, and label the document "Lab 6.2 Tableau Submission.docx".
2. Answer the questions in Connect and upload your Lab 6.2 Tableau Submission.docx via Connect, if assigned.

Lab 6.2 Power BI

Descriptive Analytics: Analyzing Sales Revenue by Product with a Tree Map

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Tree Map, New Column, New Measure, TODAY() Function

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" As discussed in the text, the type of data determines the best type of visualizations. In this lab, we analyze proportional data, with proportion defined as the number of observations in one category divided by the grand total of all observations. Using proportions is one of the primary methods of summarizing categorical data. One way to visualize proportional data is a tree map, which uses size and color to show the proportional size of values.

Sláinte is a fictional brewery that sells six different beers. It has recently expanded the distribution of its products from one state to nine states, and now its business has begun stabilizing after the expansion. With that stability comes a need for better analysis.

Assume Sláinte has hired you to help management better understand the company's sales data and provide input for its strategic decisions. Management wants to know total sales revenue for the period as well as the proportion of total sales revenue for each beer. In this lab, you will create a tree map for Sláinte using Power BI to break down total sales revenue per product sold by Sláinte. The type of beer is categorical data.

Required

1. Calculate total sales revenue.
2. Create a tree map for sales revenue broken down by beer, showing the proportion of sales revenue attributable to each beer.
3. Create a dashboard showing the tree map, total sales revenue, and date of the report (today's date). The dashboard is static, or not interactive, meaning that it does not have built-in filters to change the display. This dashboard creates a report to present the results of analytics.

Specify the Questions

1. What is total sales revenue?
2. What proportion of total sales revenue did each beer generate?

Obtain the Data

This lab uses the data in Lab 6.2 Data.xlsx, which contains sales transaction data for Sláinte from January through June of 2026.[11] The data dictionary for this file appears below. The data include three worksheets/tables: Sales_Transactions, Customers, and Products. In the Sales_Transaction worksheet/table, the data include an observation/row for each sales transaction and product sold. So, if a transaction included the sale of two products, the data set will have one unique Sales_Order_ID with two rows (that is, one for each product).

Data Dictionary

Sales_Transaction Table:

Sales_Order_ID:	Unique sales transaction identifier
Sales_Order_Date:	Date of sales order

[11] These data are a slightly altered version of Lab 3-4 data in Richardson, Terrell, and Teeter, 2021, *Introduction to Data Analytics for Accounting.* New York, NY: McGraw-Hill.

Sales_Employee_ID:	Identification number of the employee making the sale
Customer_ID:	Unique customer identifier
Product_Code:	Unique product identifier
Sales_Order_Quantity_Sold:	Number of cases sold in the sales transaction

Customers Table:

Customer_ID:	Unique customer identifier
Business Name:	Customer name
Customer_Address:	Customer street address
Customer_City:	Customer city
Customer_St:	Customer state
Customer_Zip:	Customer zip code

Products Table:

Product_Code:	Unique product identifier
Product_Description:	Type of beer
Product_Sale_Price:	Price of one case of beer

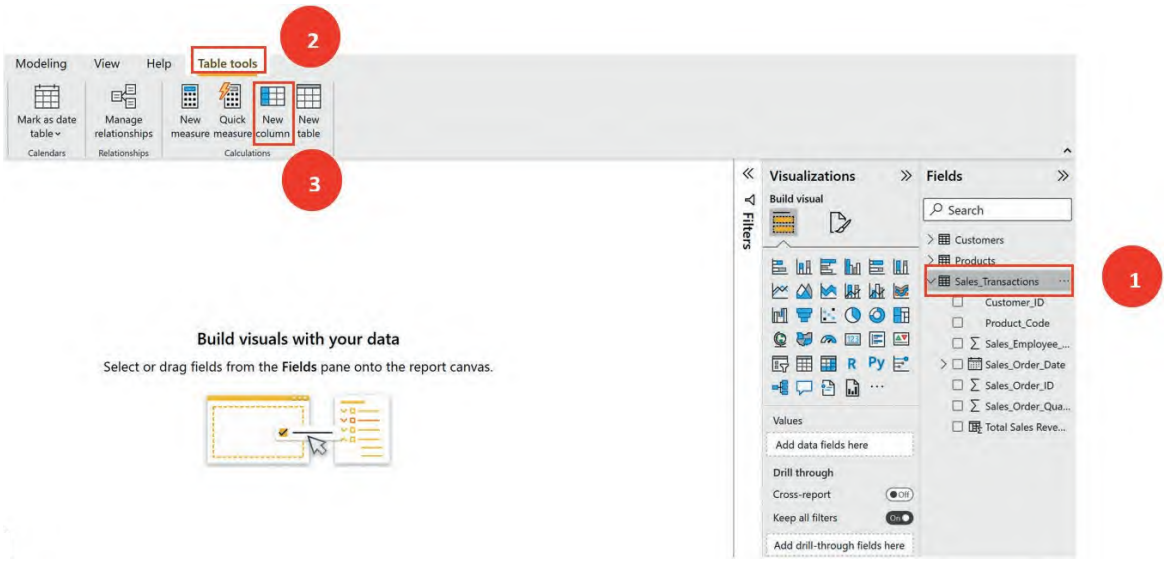
Obtain Task 1: Open Power BI and Prepare the Data.

1. Open Power BI.
2. Click Get Data, then click Connect.
3. In the Open window, browse to the Lab 6.2 Data.xlsx file.
4. Click Open.
5. Place check marks next to each table name (Customers, Products, and Sales_ Transactions) and click Load.

Analyze the Data

Analysis Task 1: Calculate Total Sales Revenue.

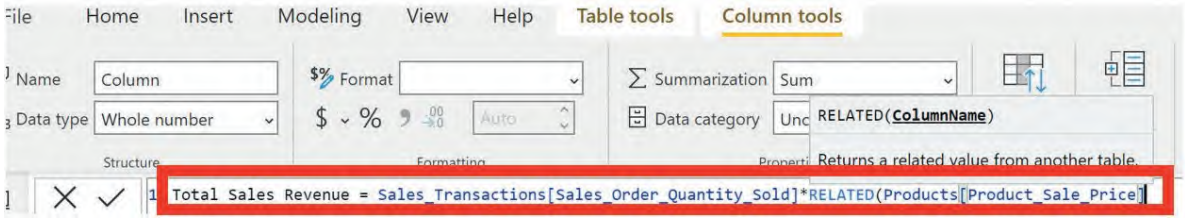
1. To determine total sales revenue, create a new column as follows: Click the Sales_ Transactions table from the Fields pane, click the Table tools tab on the ribbon, and select New column, as shown in Lab Exhibit 6.2P.1. Note, the tables names may only be showing in the Fields pane.



Lab Exhibit 6.2P.1

Microsoft Power BI

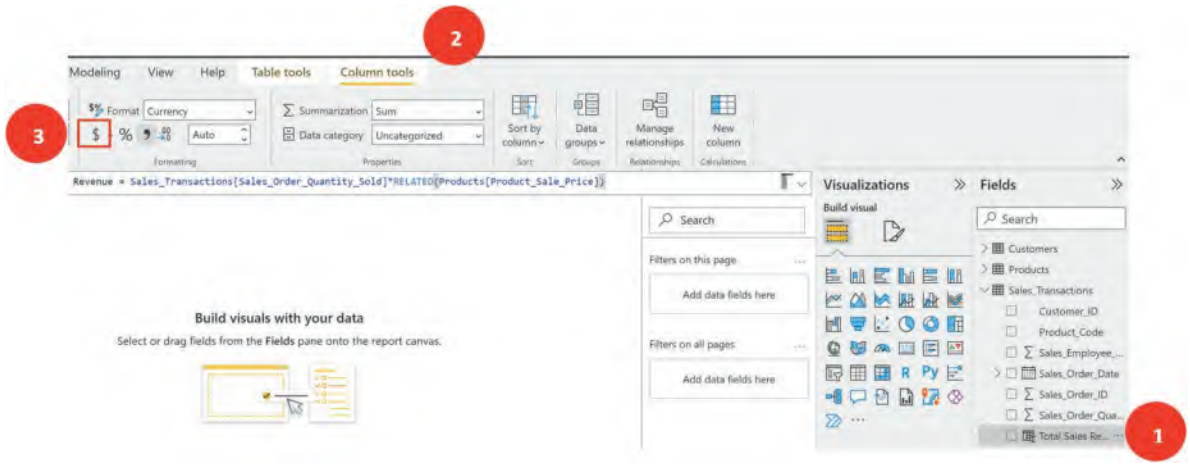
2. In the field that appears, type the following function (overwriting Column =) as shown in Lab Exhibit 6.2P.2: Total Sales Revenue = Sales_Transactions[Sales_Order_Quantity_Sold] * RELATED(Products[Product_Sale_Price]). Then click Enter.



Lab Exhibit 6.2P.2

Microsoft Power BI

3. To format total sales revenue as currency, click/highlight the field Total Sales Revenue in the Fields pane, select the Column tools tab from the ribbon, and adjust the data type to Currency, as shown in Lab Exhibit 6.2P.3.



Lab Exhibit 6.2P.3

Microsoft Power BI

Analysis Task 2: Create a Tree Map of Sales Revenue by Beer.



DATA VISUALIZATION

1. Click Treemap from the Visualizations pane with the Build visual icon selected, as shown in Lab Exhibit 6.2P.4.



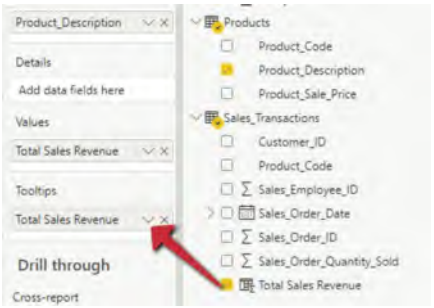
Lab Exhibit 6.2P.4

Microsoft Power BI

2. In the Fields pane, place check marks next to Total Sales Revenue and Product_ Description to fill in the tree map.
3. Power BI cannot display multiple measures in a tree map, but we can add the proportion/percentage of sales revenue by beer as a tool tip, so that we can assess the proportion when we hover over different areas of the tree map. Click on the tree map. Drag Total Sales Revenue to the Tooltips area of the Visualizations pane with the Build visual option selected, and click the drop-down arrow on the field as shown in Lab Exhibit 6.2P.5.

Lab Exhibit 6.2P.5

Microsoft Power BI



4. From the menu that appears, select **Show value as > Percent of grand total**, as shown in Lab Exhibit 6.2P.6.

Lab Exhibit 6.2P.6

Microsoft Power BI



Analysis Task 3: Label the Sheet with Today's Date as the Report Date.

When analyzing data, it is important to know the date on which the report was run. Let's use today's date as the report date.

1. From the Home tab, select New measure (in the Calculations group) and type the following into the field that appears: Today = TODAY()
   Click Enter. A new field called Today should appear in the Fields pane in the Customers Table. We will use this new measure in the report header.
2. Insert a text box to include three key items: the company name, the grand total of Sales Revenue, and Today's Date. Power BI is very intuitive with creating fields in text boxes.
   a. From the Insert tab on the ribbon, select Text box.
   b. In the text box on the canvas, type "Slainte", then click Enter to move to the next line of your text box, and type "Sales Revenue by Beer" and click Enter.
   c. On the next line, type "Total Sales Revenue" (including a space at the end), then click +Value and create a dynamic value by typing "Total Sales Revenue" in the How would you calculate this value box provided. Click the $ (Display the data as currency), and click Save. Press Enter to move to the next line of your text box.
   d. On the next line, type "Report Date " (including a space at the end), then click +Value and create a dynamic value by typing "Today" in the How would you calculate this value box provided. Adjust the date format as you wish, then click Save.
3. Format the report.
   a. Rearrange your visualizations by dragging and dropping them around your canvas to ensure the text box is in the title position (the top of the page).

b. Remove the heading above the tree map: Total Sales Revenue and %GT Total Sales Revenue by Product_Description. Click on the tree map. In the Visualizations pane, click the Format visual icon and click on the General tab. Set the Title to Off.
c. Rename the Tooltip description of %GT Total Sales Revenue. Click on the tree map. In the Visualizations pane, click the Build visual icon. Click on the dropdown v by %GT Total Sales Revenue in Tooltips. Select Rename for this visual. Type in "Percent of Total Sales".
d. Rename the Tooltip description of Total Sales Revenue. Click on the tree map. In the Visualizations pane, click the Build visual icon. Click on the dropdown v by Total Sales Revenue in Values. Select Rename for this visual. Type in "Product Total Sales".



Assessment

1. Take a screenshot of your dashboard, copy it into a Word document, and label the document "Lab 6.2 Power BI Submission.docx".
2. Answer the questions in Connect and upload your Lab 6.2 Power BI submission.docx via Connect, if assigned.

Alternate Lab 6.2

Descriptive Analytics: Analyzing Purchases by Raw Material with a Tree Map

Using either Power BI or Tableau, apply the same steps in Lab 6.2 to the Alt Lab 6.2 Data.xlsx data set. While Lab 6.2 contains data for transactions with customers, the Alt Lab 6.2 data set contains purchasing transactions with vendors/suppliers for raw materials, which will be converted into the beer that Sláinte will sell to customers. A data dictionary is presented in the Obtain the Data Section.

Specify the Questions

1. What are total purchases?
2. What proportion of total purchases is associated with each raw material?

Obtain the Data

This lab uses the data in Alt Lab 6.2 Data.xlsx, which contains supplier/vendor transaction data for Sláinte from January through June of 2026.[12] The data dictionary for this file appears below. The data include three worksheets/tables: Purchase_Orders, Suppliers, and Raw_Materials. In the Purchase_Orders worksheet/table, the data include an observation/row for each purchase transaction and type of raw material purchased. So, if a transaction included the purchase of two different raw materials, the data set will have one unique Purchase_Order_ID with two rows (that is, one for each raw material type). You will use only the Purchase Orders table and the Raw Materials table.

[12]These data are a slightly altered version of Lab 4-5 data in Richardson, Terrell, and Teeter, 2021, *Introduction to Data Analytics for Accounting.* New York, NY: McGraw Hill.

Data Dictionary

Purchase_Orders Table:

Purchase_Order_ID:	Unique purchase transaction identifier
Purchase_Order_Date:	Date of purchase order
Purchase_Employee_ID:	Unique identification number of the employee making the purchase
Supplier_ID:	Unique supplier/vendor identifier
Raw_Materials_Code:	Unique raw material product identifier
Quantity_Purchased:	Quantity of raw material purchased in the transaction

Suppliers Table:

Supplier_ID:	Unique supplier/vendor identifier
Supplier_Company_Name:	Supplier/vendor name
Supplier_Address:	Supplier/vendor street address
Supplier_City:	Supplier/vendor city
Supplier_St:	Supplier/vendor state
Supplier_Zip:	Supplier/vendor zip code

Raw_Materials Table:

Raw_Materials_Code:	Unique raw material product identifier
Raw_Materials_Description:	Description of raw material
Raw_Materials_Price:	Price of one unit of raw materials

Required

1. Calculate total purchases. Total purchases equals [Quantity Purchased]*[Raw Materials Price].
2. Create a tree map for purchases broken down by raw material showing the proportion of purchases attributable to each raw material.
3. Create a dashboard showing the tree map, total purchases, and date of the report (today's date).

Assessment

1. Take a screenshot of your dashboard, copy it into a Word document, and label the document "Alt Lab 6.2 Tableau Submission" or "Alt Lab 6.2 Power BI Submission.docx".
2. Answer the questions in Connect and upload your Alt Lab 6.2 Submission.docx via Connect, if assigned.

Lab 6.3 Tableau

Descriptive Analytics: Analyzing Sales Revenue by Customer with a Bar Chart and Filters

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Filters, Calculated Field, TODAY(), Bar Charts, Stacked Bar Chart

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" As discussed in the text, the type of data determines the best type of visualizations. In this lab, we analyze categorical data. One way to visualize categorical data is a bar chart, which uses vertical/horizontal bars with varying heights/lengths proportional to the values they represent. Comparing the height/length of the bars allows us to compare the frequency of different categories in the data set.

Sláinte is a fictional brewery that sells six different beers. It has recently expanded the distribution of its products from one state to nine states, and now its business has begun stabilizing after the expansion. With that stability comes a need for better analysis.

Assume Sláinte has hired you to help management better understand the company's sales data and provide input for its strategic decisions. Management wants to know total sales revenue for the period as well as the proportion of total sales revenue by each customer. In this lab, you will use Tableau to create a bar chart breaking down total sales revenue by customers who purchased the beer. "Customer" is an example of categorical data.

When you import data into Tableau, it automatically assigns a data type (such as names, dates, or numbers) to each field. As part of the ETL (extract, transform, and load) process, these data types need to be verified and changed if necessary. Based on the data type, Tableau assigns a role to the field: either a dimension or a measure.

- *Measures* are numeric fields that can be used for mathematical calculations.
- *Dimensions* are non-numeric and used to categorize data. Dimensions include fields identified as string (text), dates, and geographic data types.

In this lab, you will also add a drop-down filter that allows users to drill down or examine state(s) and/or specific customer(s) in more detail.

Required

1. Calculate total sales revenue.
2. Create a bar chart for total sales revenue (dollars) broken down by customer.
3. Add filters for customer state and customer business name.

Specify the Questions

1. What is total sales revenue?
2. How much sales revenue did each customer generate?
3. How much sales revenue did a state generate?
4. How much sales revenue did a product generate?

Obtain the Data

This lab uses the data in Lab 6.3 Data.xlsx, which contains sales transaction data for Sláinte from January through June of 2026.[13] The data dictionary for this file appears below. The data include three worksheets/tables: Sales_Transactions, Customers, and Products. In the Sales_Transactions worksheet/table, the data include an observation/row for each sales

[13]These data are a slightly altered version of Lab 3-4 data in Richardson, Terrell, and Teeter, 2021, *Introduction to Data Analytics for Accounting.* New York, NY: McGraw Hill.

transaction and product sold. So, if a transaction included the sale of two products, the data set will have one unique Sales_Order_ID with two rows (that is, one for each product).

Data Dictionary

Sales_Transactions Table:

Sales_Order_ID:	Unique sales transaction identifier
Sales_Order_Date:	Date of sales order
Sales_Employee_ID:	Unique identification number of the employee making the sale
Customer_ID:	Unique customer identifier
Product_Code:	Unique product identifier
Sales_Order_Quantity_Sold:	Number of cases sold in the sales transaction

Customers Table:

Customer_ID:	Unique customer identifier
Business Name:	Customer name
Customer_Address:	Customer street address
Customer_City:	Customer city
Customer_St:	Customer state
Customer_Zip:	Customer zip code

Products Table:

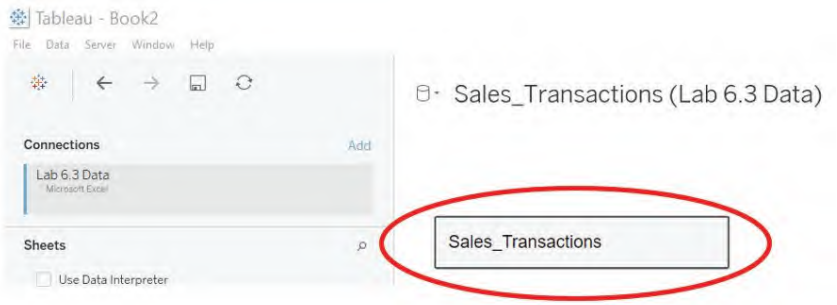
Product_Code:	Unique product identifier
Product_Description:	Type of beer
Product_Sale_Price:	Price of one case of beer

Obtain Task 1: Open Tableau and Prepare the Data.

1. Open Tableau.
2. Under Connect, select Microsoft Excel.
3. In the Open window, browse to the Lab 6.3 Data.xlsx file.
4. Click Open.
5. In Tableau 2020.2 and higher, the data source pane defaults to a logical layer for building relationships. For this lab, we need to switch the view to the physical layer to work with joins. To do so, add the Sales_Transactions sheet to the logical layer by following these directions: Drag Sales_Transactions to the data source pane, then double-click the Sales_Transactions label in the logical layer (as shown in Lab Exhibit 6.3T.1).

Lab Exhibit 6.3T.1

Tableau Software



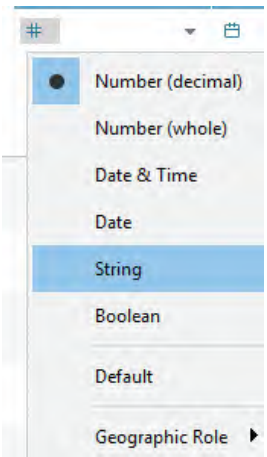
6. Drag all the remaining sheets/tables to the Data Connection page. Verify that you have inner joins created for Sales_Transactions – Customers and for Sales_ Transactions – Products, as shown in Lab Exhibit 6.3T.2.



Lab Exhibit 6.3T.2

Tableau Software

7. If Tableau imported Sales Order ID as a number, change the data type to string as it is not a numeric number for calculations. To change the type, click on the # above Sales Order ID and change it to String, as shown in Lab Exhibit 6.3T.3.



Lab Exhibit 6.3T.3

Tableau Software

8. Change Customer ID, Product Code, and Sales Employee ID to a String data type as needed.
9. If Tableau imported the Customer_St field as String, change the data type to **Geographic > State/Province,** as shown in Lab Exhibit 6.3T.4. Your data should be ready to use.

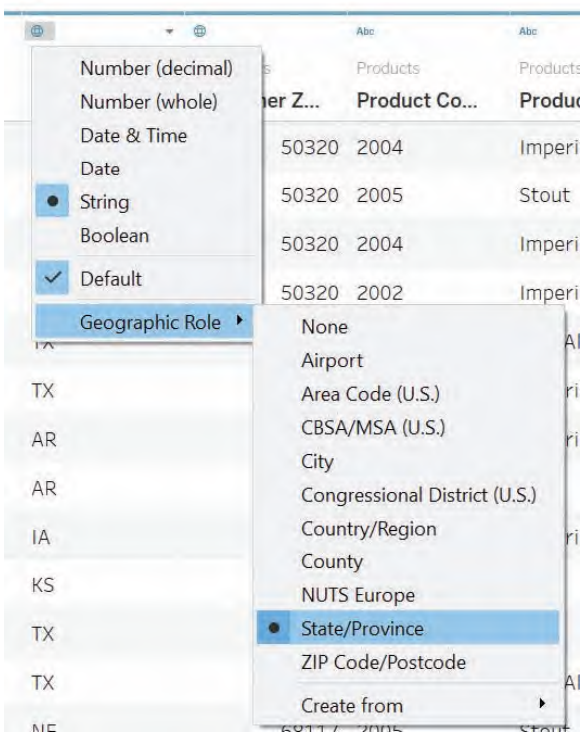
Analyze the Data

Analysis Task 1: Calculate Total Sales Revenue.

1. Click on Sheet 1.
2. To determine total sales revenue, create a calculated field as follows. Click on the down arrow at the top of the Data tab. Select Create Calculated Field..., as shown in Lab Exhibit 6.3T.5.

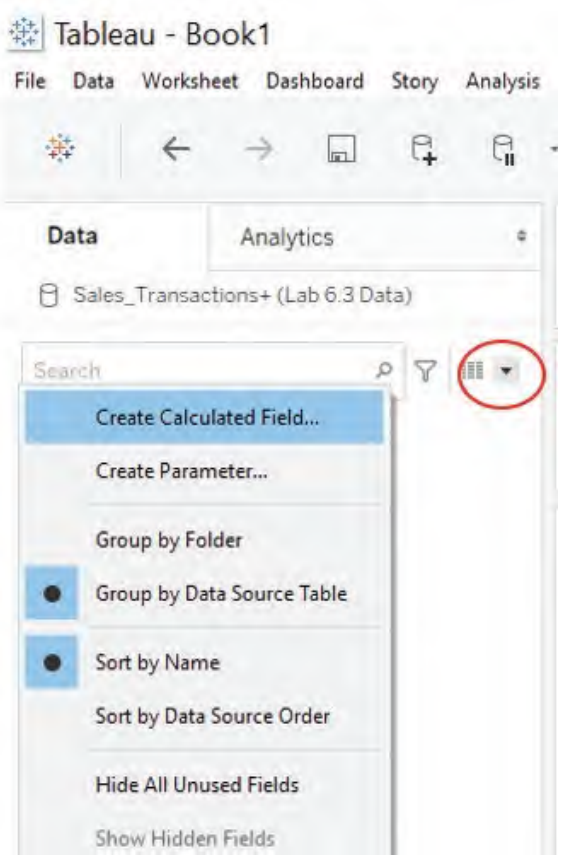
Lab Exhibit 6.3T.4

Tableau Software



Lab Exhibit 6.3T.5

Tableau Software



3. In the dialog box that appears, type "Total Sales Revenue" (over the grayed Calculation1) as the variable name. Type the formula for total sales, which equals number of units sold times sales price, or Sales Order Quantity Sold * Product Sale Price. As you type in the variable names, Tableau shows the matching variables; select the appropriate variable name. Lab Exhibit 6.3T.6 shows the dialog box with the formula.



Lab Exhibit 6.3T.6

Tableau Software

4. Verify that the bottom-left corner of the box shows "The calculation is valid." If so, click OK. If not, correct any errors. The new variable, Total Sales Revenue, should appear in the Data tab towards the bottom with the Measures.

Analysis Task 2: Create the Bar Chart.

1. Drag Customer ID from Sales_Transactions to the Columns shelf.
2. Drag Total Sales Revenue to the Rows shelf.
3. To quickly identify the customer with the largest amount of sales revenue, sort the data in descending order using the sort icon, as shown in Lab Exhibit 6.3T.7.



Lab Exhibit 6.3T.7

Tableau Software

4. Now edit the Y-axis label to clearly show that it is in dollars. Right-click the Y-axis. Select Edit Axis. Change the Axis Title to Total Sales Revenue (Dollars). Click x to close.

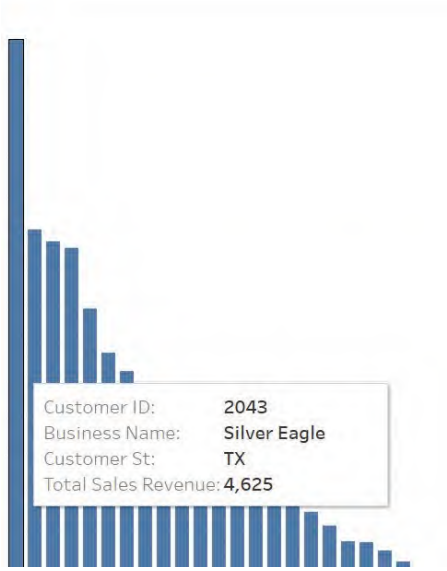
Analysis Task 3: Format the Visualization Tooltip.

When analyzing customer data, it is helpful to know the customer name, not just the customer ID, which is currently shown on the visualization. There are too many customers to put the customers' business names on the visualization, so let's add business name to the tooltip. The *tooltip* shows information when the cursor is placed over the column. In addition, let's also have the tooltip show customer state, which may be helpful in decision-making.

1. Drag Business Name to Tooltip in the Marks card.
2. Drag Customer St to Tooltip in the Marks card. Now, when the cursor is placed on a column, the Customer ID, Business Name, Customer St, and Total Sales Revenue appear, as shown in Lab Exhibit 6.3T.8.

Lab Exhibit 6.3T.8

Tableau Software



3. Let's format the Total Sales Revenue as currency. Click the down arrow by Total Sales Revenue in the Data pane. Select **Default Properties > Number format. . .** The Default Number Format [Total Sales Revenue] dialog box should appear. Select Currency (Custom) and set decimals to 0. Click OK. Now when you use the Tooltip, Total Sales Revenue should be formatted as currency. For example, 4,625 should appear now as $4,625.

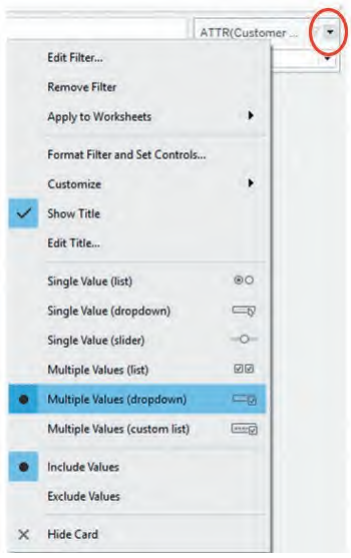
Analysis Task 4: Add Drop-Down Filters.

Another way to help users analyze the data is to add filters. Let's add drop-down sheet filters for Customer St and Business Name.

1. From the menu, select **Analysis > Filters > Customer St.**
2. Change the list to a drop-down. Click on the down arrow by ATTR(Customer St) at the top-right corner of the canvas. Select Multiple Values (dropdown), as shown in Lab Exhibit 6.3T.9. The list should change to a drop-down that allows users to analyze sales revenue for all, one, or a group of states. If Show Me is in the way, click on Show Me to make it go to the back.

Lab Exhibit 6.3T.9

Tableau Software



3. Add a drop-down filter for Business Name using the process described above.
4. To make the data visualization more visually appealing, you can change the drop-down labels. For example, change the drop-down filter name for Customer St to just State. Click on the down arrow by ATTR(Customer St) at the top-right corner of the canvas. Select Edit Title. Type "State". Click OK.

Analysis Task 5: Add a Title with Today's Date.

When analyzing data, it is important to know the date on which the report was run. Let's use today's date as the report date.

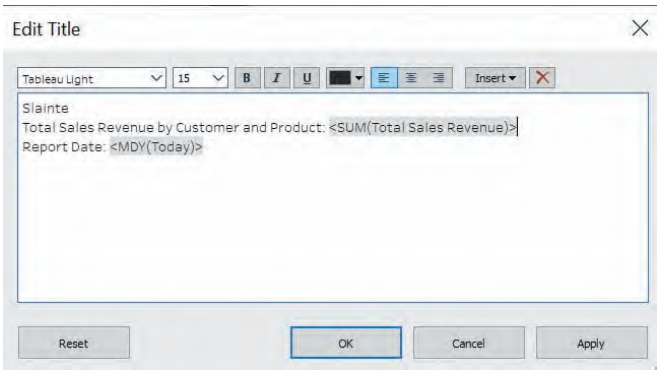
1. Create a new calculated field called Today with the formula equal to TODAY().
2. Drag Today onto Detail in the Marks card. Right-click YEAR(Today). Select **More > Custom.** In the Custom Date box, use the drop-down arrow to select Month/Day/Year, as shown in Lab Exhibit 6.3T.10. Click OK. The current date will now show in the Tooltip and is available to use in the sheet title in Step 3 below.



Lab Exhibit 6.3T.10

Tableau Software

3. Let's create a sheet title. Right-click on Sheet 1 at the top of the canvas. Select Edit Title.
4. In the Edit Title dialog box, on the first line type "Slainte".
5. On the second line, type "Sales Revenue by Customer". Click on the down arrow by Insert. Select SUM(Total Sales Revenue).
6. On the third line, type "Report Date:". Click on the down arrow by Insert. Select MDY(Today). The Edit Title Box should look like Lab Exhibit 6.3T.11. Click OK. The sheet should now have a title with the report date, which will change with the current date.



Lab Exhibit 6.3T.11

Tableau Software

7. Name the Sheet 1 tab “Sales Revenue by Customer”.

DATA VISUALIZATION

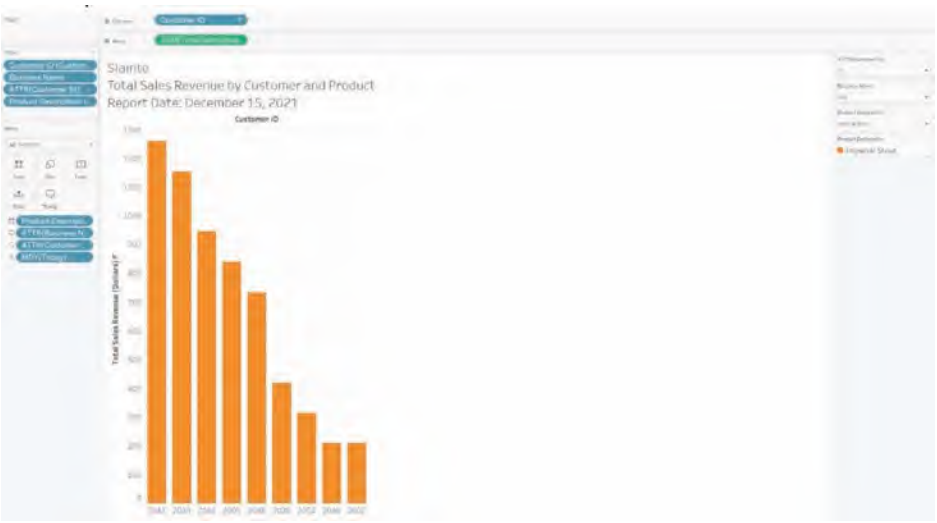
Report the Results: Add Color and a Filter for Product

You just created a bar chart with two drop-down filters. Assume that management wants the breakdown by beer as well. So, let’s add color to the bar chart for beer and a drop-down filter that allows the user to also select one or more beers.

1. Drag Product Description to Color in the Marks Card. Each column should be broken down into colors showing different types of beer that were sold. A stacked bar chart has been created. A legend defining the colors should automatically appear to the right underneath the drop-down filters.
2. The graph may seem a little colorful now, so let’s add a drop-down filter for product. Following the same procedure described for Customer St above, add a drop-down filter for Product Description using the **Analysis > Filters > Product Description** menu option. The Product Description legend should still appear and changes based on the selected products.
3. Reorder the filters as you choose by dragging and placing the drop-down filter to the desired location.
4. Let’s see how the drop-down filters work. Select all customers from Texas that purchase Imperial Stout beer. Lab Exhibit 6.3T.12 shows the results of this selection. Note that the report date will be the current date.

Lab Exhibit 6.3T.12

Tableau Software



5. Notice that the bottom-left corner says: “SUM(Total Sales Revenue): 6,090”. This indicates the total dollar amount of sales revenue for Imperial Stout purchased by Texas customers.

Assessment

1. Take a screenshot of your Tableau worksheet with the filters in place, copy it into a Word document named “Lab 6.3 Tableau Submission.docx”, and label the screenshot Submission 1.
2. Clear the filters so that all observations are showing. Take a screenshot of your worksheet, copy it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 6.3 Tableau Submission.docx via Connect, if assigned.

Lab 6.3 Power BI

Descriptive Analytics: Analyzing Sales Revenue by Customer with a Bar Chart and Filters

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Filters, Calculated Field, TODAY(), Bar Charts, Stacked Bar Chart

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" As discussed in the text, the type of data determines the best type of visualizations. In this lab, we analyze categorical data. One way to visualize categorical data is a bar chart, which uses vertical/horizontal bars with varying heights/lengths proportional to the values they represent. Comparing the height/length of the bars allows us to compare the frequency of different categories in the data set.

Sláinte is a fictional brewery that sells six different beers. It has recently expanded the distribution of its products from one state to nine states, and now its business has begun stabilizing after the expansion. With that stability comes a need for better analysis.

Assume Sláinte has hired you to help management better understand the company's sales data and provide input for its strategic decisions. Management wants to know total sales revenue for the period as well as the proportion of total sales revenue by each customer. In this lab, you will use Power BI to create a bar chart breaking down total sales revenue by customers who purchased the beer. "Customer" is an example of categorical data.

When you import data into Power BI, it automatically assigns a data type (such as names, dates, or numbers) to each field. As part of the ETL (extract, transform, and load) process, these data types need to be verified and changed if necessary. Based on the data type, Power BI assigns a role to the field: either a dimension or a measure.

- *Measures* are numeric fields that can be used for mathematical calculations.
- *Dimensions* are non-numeric and used to categorize data. Dimensions include fields identified as string (text), dates, and geographic data types.

In this lab, you will also add a drop-down filter that allows users to drill down or examine state(s) and/or specific customer(s) in more detail.

Required

1. Calculate total sales revenue.
2. Create a bar chart for total sales revenue (dollars) broken down by customer.
3. Add filters for customer state and customer business name.

Specify the Questions

1. What is total sales revenue?
2. How much sales revenue did each customer generate?
3. How much sales revenue did a state generate?
4. How much sales revenue did a product generate?

Obtain the Data

This lab uses the data in Lab 6.3 Data.xlsx, which contains sales transaction data for Sláinte from January through June of 2026.[14] The data dictionary for this file follows.

[14]These data are a slightly altered version of Lab 3-4 data in Richardson, Terrell, and Teeter, 2021, *Introduction to Data Analytics for Accounting.* New York, NY: McGraw Hill.

The data include three worksheets/tables: Sales_Transactions, Customers, and Products. In the Sales_Transaction worksheet/table, the data include an observation/row for each sales transaction and product sold. So, if a transaction included the sale of two products, the data set will have one unique Sales_Order_ID with two rows (that is, one for each product).

Data Dictionary

Sales_Transaction Table:

Sales_Order_ID:	Unique sales transaction identifier
Sales_Order_Date:	Date of sales order
Sales_Employee_ID:	Identification number of the employee making the sale
Customer_ID:	Unique customer identifier
Product_Code:	Unique product identifier
Sales_Order_Quantity_Sold:	Number of cases sold in the sales transaction

Customers Table:

Customer_ID:	Unique customer identifier
Business Name:	Customer name
Customer_Address:	Customer street address
Customer_City:	Customer city
Customer_St:	Customer state
Customer_Zip:	Customer zip code

Products Table:

Product_Code:	Unique product identifier
Product_Description:	Description of beer
Product_Sale_Price:	Price of one case of beer

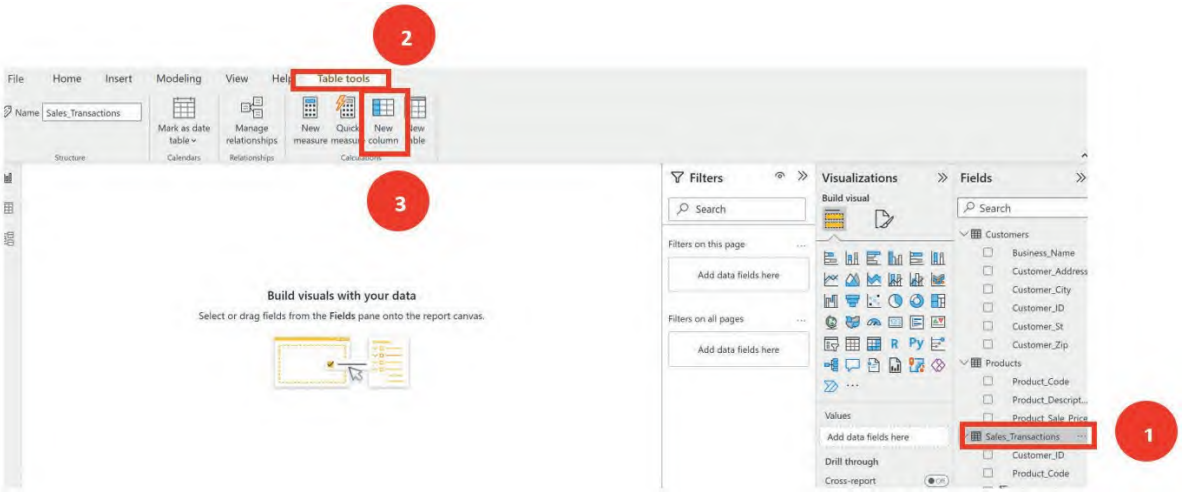
Obtain Task 1: Open Power BI and Prepare the Data.

1. Open Power BI.
2. Click Get Data, and then click Connect.
3. In the Open window, browse to the Lab 6.3 Data.xlsx file.
4. Click Open.
5. Place check marks next to each table name (Customers, Products, and Sales_Transactions) and click Load.
6. Power BI imported Customer_ID as a Whole number, but it should be Text. In the Fields pane, highlight Customer_ID in the Customers table. The Column tools tab should appear. Change Data type to Text. Repeat the process for the Customer_ID, Product_Code, Sales_Employee_ID, and Sales_Order_ID in the Sales_Transactions table as well as for Product_Code and Product_Description (if necessary) in the Products table.

Analyze the Data

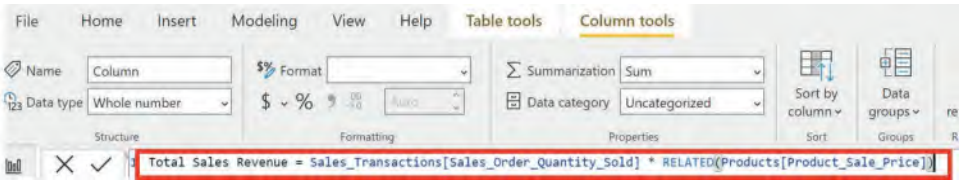
Analysis Task 1: Calculate Total Sales Revenue.

1. To determine total sales revenue, create a new column as follows: Click the Sales_Transactions table from the Fields pane, click the Table tools tab on the ribbon, and select New column, as shown in Lab Exhibit 6.3P.1.
2. In the field that appears, type the following function as shown in Lab Exhibit 6.3P.2: Total Sales Revenue = Sales_Transactions[Sales_Order_Quantity_Sold] * RELATED(Products[Product_Sale_Price]) and click Enter.



Lab Exhibit 6.3P.1

Microsoft Power BI



Lab Exhibit 6.3P.2

Microsoft Power BI

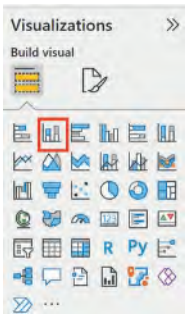
3. The new column, Total Sales Revenue, should appear in the fields beneath Sales_ Transactions in the Fields pane. To format the new column as currency, click the $ in the Column tools tab.

Analysis Task 2: Create the Bar Chart.

1. Click Stacked Column Chart from the Visualizations pane with the Build visual icon selected, as shown in Lab Exhibit 6.3P.3.



DATA VISUALIZATION



Lab Exhibit 6.3P.3

Microsoft Power BI

2. Place check marks next to Total Sales Revenue and Customer_ID to fill in the column chart.
3. If the stacked column chart is not sorted in descending order, click on the stacked column chart. Then click on the . . .Select **Sort axis >,** check Total Sales Revenue, and check Sort descending.
4. Edit the Y-axis label to clearly show that it is in dollars. In the Visualizations pane, click on the Format visual icon, verify you are in the Visual tab, and expand the Y-axis to see customization options. Expand the Title Option, and replace the Title text of Auto with "Total Sales Revenue (Dollars)".
5. Change the X-axis label to "Customer ID".
6. In the Visualizations pane, click the Format visual icon, click General, and set Title to Off.

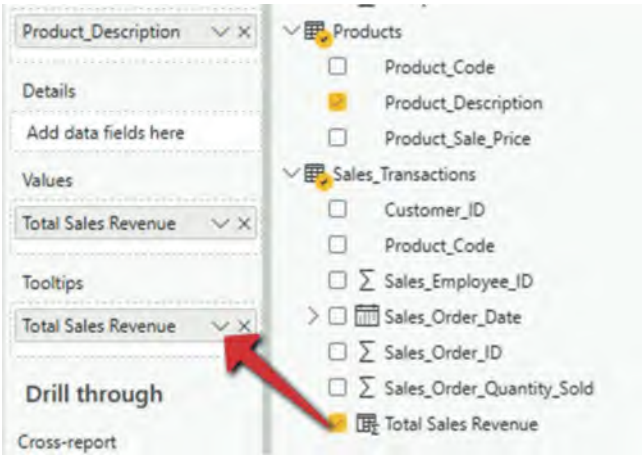
Analysis Task 3: Format the Visualization Tooltip.

Power BI cannot display multiple measures in a column chart, but we can add the proportion/percentage of sales revenue by beer as a tooltip. The *tooltip* shows information when the cursor is placed (or hovers) over a column.

1. In the Visualizations pane, click on the Build visual icon. Drag Total Sales Revenue to the Tooltips Add data fields here box, and click the drop-down arrow, as shown in Lab Exhibit 6.3P.4.

Lab Exhibit 6.3P.4

Microsoft Power BI



2. From the menu that appears, select **Show value as > Percent of grand total**, as shown in Lab Exhibit 6.3P.5.

Lab Exhibit 6.3P.5

Microsoft Power BI

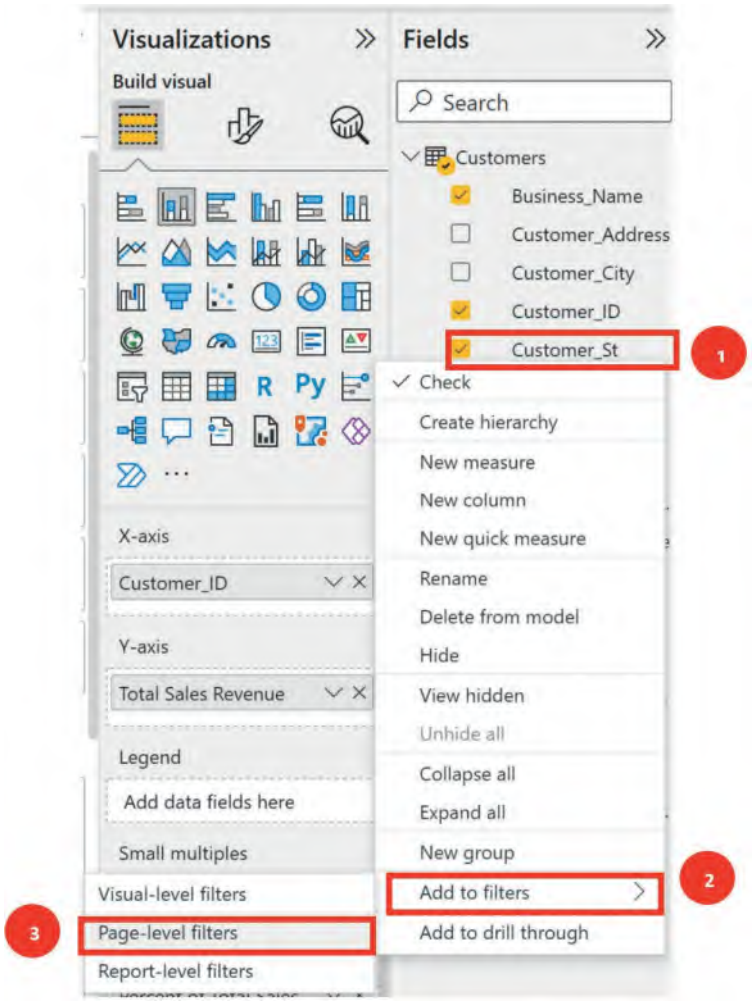


3. Let's rename the tooltip from %GT Total Sales Revenue to Percent of Total Sales Revenue. Click the down arrow again in the Tooltips box. Select Rename for this visual. Type "Percent of Total Sales" Click Enter.
4. When analyzing customer data, it is helpful to know the customer name, not just the customer ID, which is currently shown on the visualization. There are too many customers to put the customers' business names on the visualization, so let's add business name to the tooltip. In addition, let's also have the tooltip show customer state, which may be helpful in decision-making. Drag Business_Name and Customer_St to the Tooltips area below Percent of Total Sales.

Analysis Task 4: Add Filters.

Another way to help users analyze the data is to add filters. Let's add drop-down page filters for Customer St and Business Name.

1. Right-click the Customer_St field in the Fields pane, select Add to filters, and then select Page-level filters, as shown in Lab Exhibit 6.3P.6.



Lab Exhibit 6.3P.6

Microsoft Power BI

2. Add a drop-down filter for Business_Name using the same process described above.
3. Change the drop-down filter name for Customer St to just State. In the Filters pane, right-click on the filter label for Customer_St and select Rename. Then type "State" (Lab Exhibit 6.3P.7). Click Enter.
4. Change the drop-down filter name for Business_Name to just Business Name. In Power BI, the filters remain in the Filters pane. All selections are made in that pane.

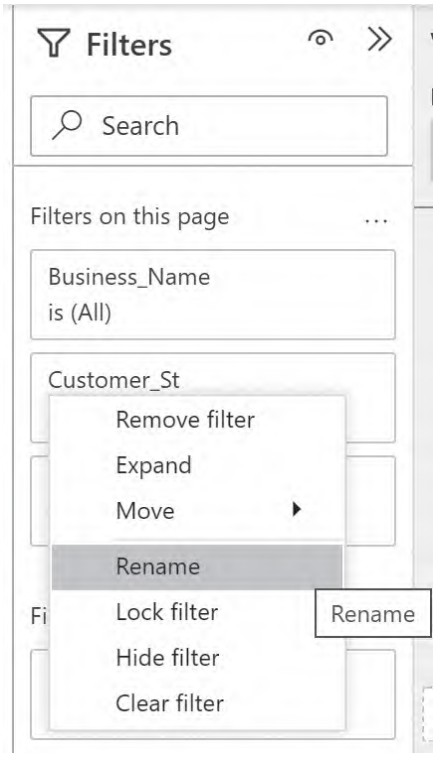
Analysis Task 5: Label the Page with Today's Date as the Report Date.

When analyzing data, it is important to know the date on which the report was run. Let's use today's date as the report date.

1. From the Home tab, select New measure and type the following into the field that appears: Today = TODAY(). Click Enter. The new measure called Today should appear in the Fields pane in the Customers table. We will use this new measure in the report header.

Lab Exhibit 6.3P.7

Microsoft Power BI

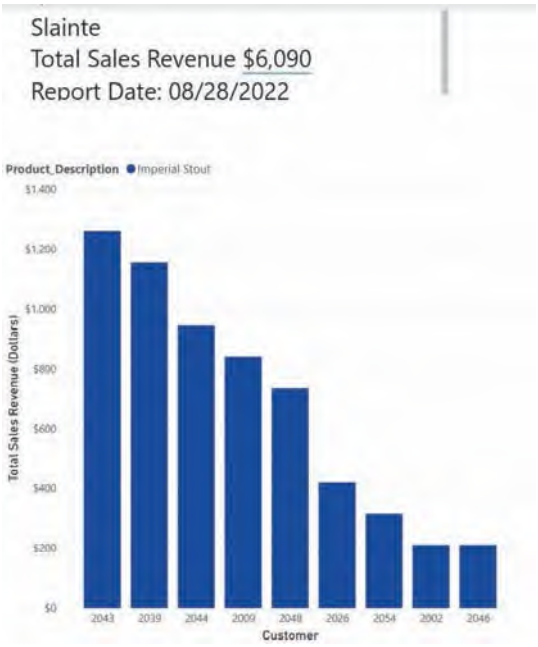


2. Insert a text box to include three key items: the company name, the grand Total Sales Revenue, and Today's Date. Power BI is very intuitive with creating fields in text boxes. From the Insert tab on the ribbon, select Text box.
   a. In the new Text box, type "Slainte", then click Enter to move to the next line of your text box.
   b. On the next line, type "Total Sales Revenue" (with a space at the end), then click +Value and create a dynamic value by typing "Total Sales Revenue" in the "How would you calculate this value" box Click the $ to adjust the data format to currency and click Save. Click Enter to move to the next line of your text box.
   c. On the next line, type "Report Date:" (with a space at the end), then click +Value and create a dynamic value by typing "Today" in the "How would you calculate this value" box. Adjust the date format as you wish, then click Save.
3. Rearrange your visualizations by dragging and dropping them around your canvas to ensure the text box is in the title position (the top of the page).

Report the Results: Add Color and a Filter for Product

You just created a bar chart with two filters. Assume that management wants the breakdown by beer as well. So, let's add color to the bar chart for beer and a drop-down filter that allows the user to also select one or more beers.

1. Click on the column chart.
2. In the Visualizations pane, click on the Build visual icon.
3. Drag Product_Description to the Legend box. A stacked bar chart has been created. A legend defining the colors should automatically appear.
4. The graph may seem a little colorful now, so let's add a filter for product. Following the same procedure described for Customer St above, add a filter for Product Description.
5. Change the X-axis Customer_ID label to just Customer. In the Visualizations pane, click the Format visual icon. Then expand X-axis, expand Title, and change Title Text to "Customer". Click Enter.
6. Let's see how the drop-down filters work. Select all customers from Texas that purchase Imperial Stout beer. Lab Exhibit 6.3P.8 shows the results of this selection.



Lab Exhibit 6.3P.8

Microsoft Power BI

Assessment

1. Take a screenshot of your Power BI worksheet with the filters in place, copy it into a Word document named "Lab 6.3 Power BI Submission.docx", and label the screenshot Submission 1.
2. Clear the filters so that all observations are showing. Take a screenshot of your worksheet, copy it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 6.3 Power BI Submission.docx via Connect, if assigned.

Alternate Lab 6.3

Descriptive Analytics: Analyzing Purchases by Supplier with a Bar Chart and Filters

Using either Power BI or Tableau, apply the same steps in Lab 6.3 to the Alt Lab 6.3 Data.xlsx data set. While Lab 6.3 contains data for transactions with customers, the Alt Lab 6.3 Data contain purchasing transactions with vendors/suppliers for raw materials, which will be converted into the beer that Sláinte will sell to customers. A new data dictionary is presented in the Obtain the Data Section.

Specify the Questions

1. What are total purchases?
2. What was the total dollar amount of raw materials purchased from each supplier/vendor?
3. What was the total dollar amount purchased of each raw material?

Obtain the Data

This lab uses the data in Alt Lab 6.3 Data.xlsx, which contains supplier/vendor transaction data for Sláinte from January through June of 2026.[15] The data dictionary for this file appears below. The data include three worksheets/tables: Purchase_Orders, Suppliers, and

[15]These data are a slightly altered version of Lab 4-5 data in Richardson, Terrell, and Teeter, 2021, *Introduction to Data Analytics for Accounting.* New York, NY: McGraw Hill.

Raw_Materials. In the Purchase_Orders worksheet/table, the data include an observation/row for each purchase transaction and type of raw material purchased. So, if a transaction included the purchase of two different raw materials, the data set will have one unique Purchase_Order_ID with two rows (that is, one for each raw material type). You will use all three tables.

Data Dictionary

Purchase_Orders Table:

Purchase_Order_ID:	Unique purchase transaction identifier
Purchase_Order_Date:	Date of purchase order
Purchase_Employee_ID:	Unique identification number of the employee making the purchase
Supplier_ID:	Unique supplier/vendor identifier
Raw_Materials_Code:	Unique raw material product identifier
Quantity_Purchased:	Quantity of raw material purchased in the transaction

Suppliers Table:

Supplier_ID:	Unique supplier/vendor identifier
Supplier_Company_Name:	Supplier/vendor name
Supplier_Address:	Supplier/vendor street address
Supplier_City:	Supplier/vendor city
Supplier_St:	Supplier/vendor state
Supplier_Zip:	Supplier/vendor zip code

Raw_Materials Table:

Raw_Materials_Code:	Unique raw material product identifier
Raw_Materials_Description:	Description of raw material
Raw_Materials_Price:	Price of one unit of raw materials

Required

1. Calculate total purchases in dollars. Total purchases in dollars equals [Quantity Purchased]*[Raw Materials Price]. Total Purchases should be formatted as currency with zero decimals.
2. Create a bar chart for total purchases dollars broken down by supplier. If you are using Tableau, change Standard view to Fit Width to see the complete Supplier Company Name.
3. Create a title with today's date.
4. Color-code each bar for raw material type.
5. Add drop-down filters for Raw Materials Description and Supplier Company Name.

Assessment

1. Take a screenshot of your dashboard, copy it into a Word document, and label the document "Alt Lab 6.3 Tableau Submission" or "Alt Lab 6.3 Power BI Submission.docx".
2. Answer the questions in Connect and upload your Alt Lab 6.3 Submission.docx via Connect, if assigned.

Lab 6.4 Excel

Descriptive Analytics: Creating a Dashboard Using Pivot Tables and Slicers

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Pivot Table, Pivot Chart, Dashboard, Slicer

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" An effective way to present multiple (descriptive) data visualizations is an interactive dashboard that presents a collection of useful visualizations, measures, and tables that allow a company to monitor progress over a period of time or at a particular point in time. An interactive dashboard allows users to dig deep into the data by filtering one or more of the visualizations in the dashboard.

Sláinte is a fictional brewery that sells six different beers. It has recently expanded the distribution of its products from one state to nine states, and now its business has begun stabilizing after the expansion. With that stability comes a need for better analysis.

Assume Sláinte has hired you to help management better understand the company's data and provide input for its strategic decisions. Management wants all of the analysis placed in an interactive dashboard.

In this lab, you will create a dashboard that shows: (1) the total cases sold of each beer with a summary bar chart, (2) the count of sales orders for each beer with a summary bar chart, and (3) the total cases sold and count of sales orders for each beer over time with a summary line chart. In addition, Sláinte wants the following two slicers on the dashboard: (1) a dashboard-level slicer to filter each pivot table and pivot chart based on Customer state and (2) a report-level slicer to filter only the sales-of-beer-over-time pivot table and pivot chart by beer.

Required

1. Create three pivot tables and pivot charts.
   a. Pivot table 1 and bar chart: Sum of Cases/Quantity Sold for each beer
   b. Pivot table 2 and bar chart: Count of Sales Orders for each beer
   c. Pivot table 3 and line chart: Sum of Cases/Quantity Sold and Count of Sales Orders for each beer over time
2. Create two slicers.
   a. A dashboard-level slicer to filter each pivot table and pivot chart based on Customer state
   b. A report-level slicer to filter only the sales-of-beer-over-time pivot table and pivot chart by beer
3. Work with the slicers to present different versions of the reports in the dashboard.

Specify the Questions

1. How many cases of each beer were sold to customers, and how many sales transactions took place in each state?
2. How many cases of each beer were sold to customers, and how many sales transactions took place in each state on each transaction date?

Obtain the Data

This lab uses the data in Lab 6.4 Data.xlsx, which contains sales transaction data for Sláinte from January through June of 2026.[16] The data dictionary for this file is below. The data

[16]These data are a slightly altered version of Lab 3-4 data in Richardson, Terrell, and Teeter. 2021. *Introduction to Data Analytics for Accounting.* New York, NY: McGraw Hill.

include three worksheets/tables: Sales_Transactions, Customers, and Products. In the Sales_Transactions worksheet/table, the data include an observation/row for each sales transaction and product sold. So, if a transaction included the sale of two products, the data set will have one unique Sales_Order_ID with two rows (that is, one for each product).

Data Dictionary

Sales_Transactions Table:

Sales_Order_ID:	Unique sales transaction identifier
Sales_Order_Date:	Date of sales order
Sales_Employee_ID:	Unique identification number of the employee making the sale
Customer_ID:	Unique customer identifier
Product_Code:	Unique product identifier
Sales_Order_Quantity_Sold:	Number of cases sold in the sales transaction

Customers Table:

Customer_ID:	Unique customer identifier
Business Name:	Customer name
Customer_Address:	Customer street address
Customer_City:	Customer city
Customer_St:	Customer state
Customer_Zip:	Customer zip code

Products Table:

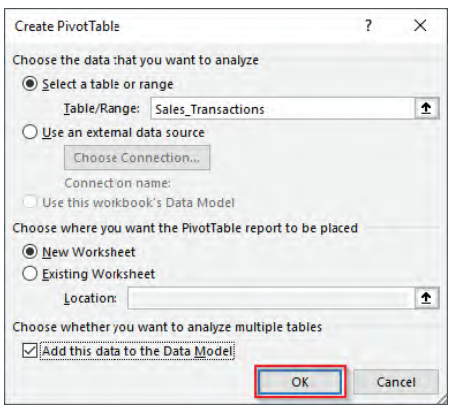
Product_Code:	Unique product identifier
Product_Description:	Description of beer
Product_Sale_Price:	Price of one case of beer

Analyze the Data

Analysis Task 1: Create the First Pivot Table.

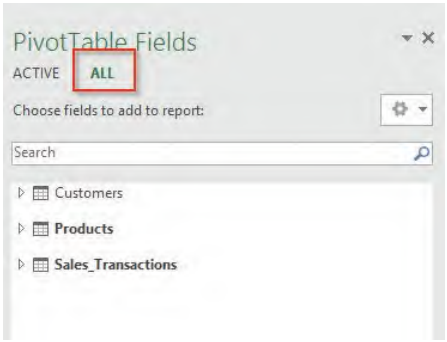
1. Insert a pivot table from any of the tables in the Excel workbook. To do this, place the cursor inside the data on any worksheet. From the menu, select **Insert > Tables > Pivot Table.** Check the box next to Add this data to the Data Model, as shown in Lab Exhibit 6.4.1.

Lab Exhibit 6.4.1



2. In the PivotTable Fields list, click ALL, as shown in Lab Exhibit 6.4.2 to show the fields from each of the three tables in the workbook and click on the arrows to see the fields. The relationships between the tables have already been created in this data set.

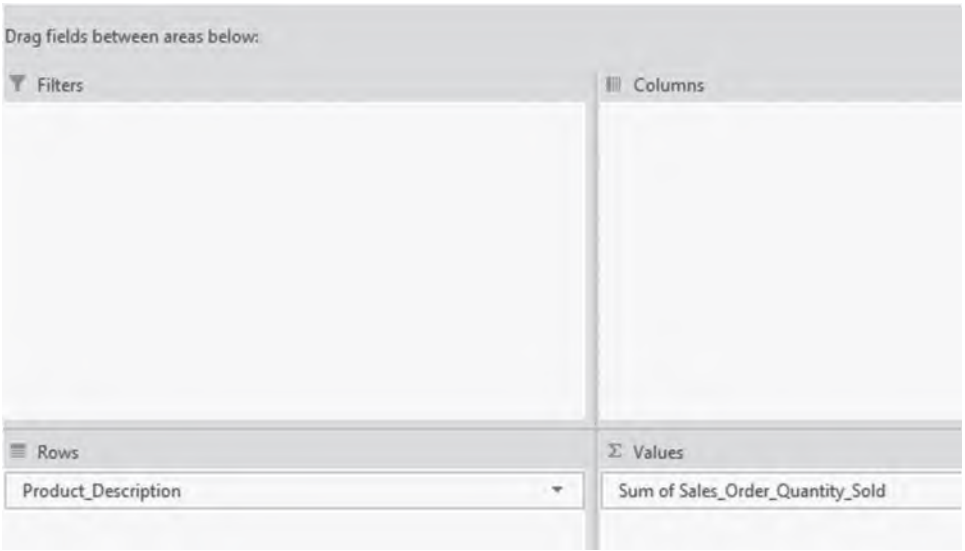
Lab Exhibit 6.4.2



3. Place the following fields in your pivot table (as shown in Lab Exhibit 6.4.3):

   - Sales_Order_Quantity_Sold (from Sales_Transactions) in Values
   - Product_Description (from Products) in Rows

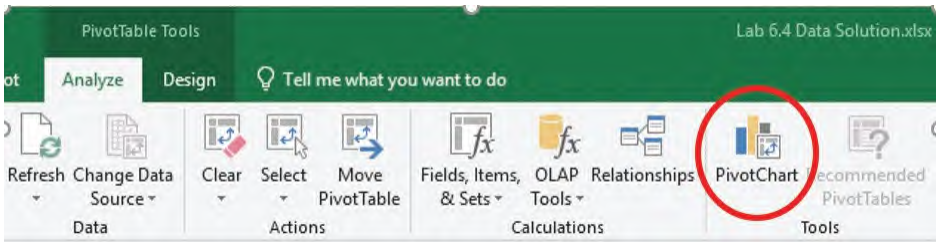
Lab Exhibit 6.4.3



Analysis Task 2: Create the First Pivot Chart.

1. Place your cursor in the pivot table. From the PivotTable Tools menu, select **Analyze > Tools > PivotChart,** as shown in Lab Exhibit 6.4.4.

Lab Exhibit 6.4.4



2. Click OK to insert the (Clustered) Column Chart (the default). A PivotChart should appear.
3. Rename the chart "Total Cases Sold by Product".
4. Add a Y-axis label. In the PivotChart Tools menu, select **Design > Chart Layouts > Add Chart Element > Axis Titles > Primary Vertical.** Change the Axis Title to "Number of Cases".
5. Add an X-axis label. In the PivotChart Tools menu, select **Design > Chart Layouts > Add Chart Element > Axis Titles > Primary Horizontal.** Change the Axis Title to "Beer".
6. To make the chart easier to read, you may want to rename the Sum Values labels. Click on the down arrow by Sum of Sales_Order_Quantity_Sold (in Sum Values). Select Value Field Settings. Change Custom Name to "Cases Sold".

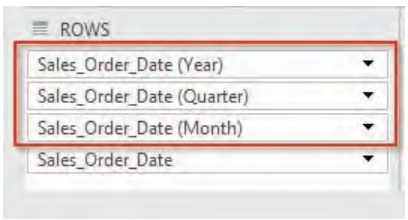
Analysis Task 3: Create the Second Pivot Table and Pivot Chart.

1. To create a second pivot table, copy the table you just created as follows: Make sure that you select all the fields in the pivot table, then press Ctrl+C on your keyboard (or right-click the selection with your mouse and select copy). Paste the copy of the existing pivot table a few rows beneath the existing one.
2. In this new pivot table, edit the field in the Values section to reflect a Count of Sales Order ID as follows: Drag Cases Sold (the renamed Sum of Sales_Order_Quantity) from Values. Drag Sales Order ID (from Sales_Transactions) to Values.
3. If Sum of Sales_Order_ID is appearing, click the down arrow to the right of the Sum of Sales_Order_ID. Select Value Field Settings. In the Value Field Settings dialog box, select Count. Click OK. Count of Sales_Order_ID will provide the total number of sales transactions.
4. Add a pivot chart for this pivot table and click OK to create the default Column Chart, as described in Analysis Task 2.
5. Rename the chart "Number of Sales Orders by Product".
6. Add a Y-axis label. In the PivotChart Tools menu, select **Design > Chart Layouts > Add Chart Element > Axis Titles > Primary Vertical.** Change the Axis Title to "Count".
7. Add an X-axis label. In the PivotChart Tools menu, select **Design > Chart Layouts > Add Chart Element > Axis Titles > Primary Horizontal.** Change the Axis Title to "Beer".
8. To make the chart easier to read, you may want to rename the Sum Values labels. Click on the down arrow by Count of Sales_Order_ID (in Sum Values). Select Value Field Settings. Change Custom Name to "Sales Orders".

Analysis Task 4: Create the Third Pivot Table with a Line Chart.

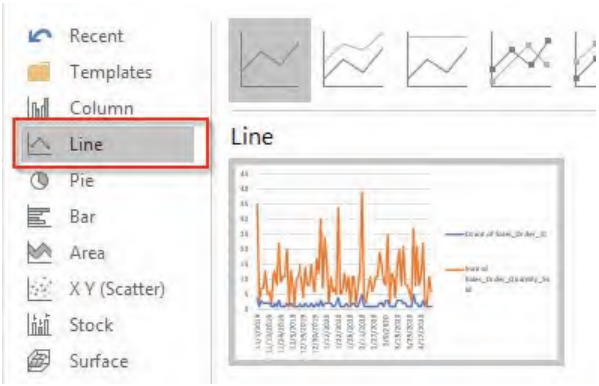
1. Copy one of your existing pivot tables, as described in Analysis Task 3.
2. Paste the new pivot table a few records below the existing pivot tables.
3. Use the following fields, removing any fields that don't match:
   - Sales_Order_Date in Rows
   - Sales_Order_ID in Values (Count)
   - Sales_Order_Quantity_Sold in Values (Sum)
4. Depending on your version of Excel, it may have automatically created a date hierarchy in the rows. If that's the case, remove all fields in the red box (Sales_Order_Date (Year), Sales_Order_Date (Quarter), and Sales_Order_Date (Month)) except for Sales_Order_Date, as shown in Lab Exhibit 6.4.5.

Lab Exhibit 6.4.5



5. Insert a Line chart for this pivot table, as shown in Lab Exhibit 6.4.6. You may need to override the default in the PivotChart window. Click OK.

Lab Exhibit 6.4.6

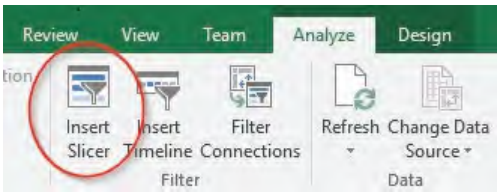


6. To make the chart easier to read, you may want to rename the Sum Values labels. For example, click on the down arrow by Count of Sales_Order_ID (in Sum Values). Select Value Field Settings. Change Count of Sales_Order_ID to "Sales Orders".
7. Add a Y-axis label. In the PivotChart Tools menu, select **Design > Chart Layouts > Add Chart Element > Axis Titles > Primary Vertical.** Change the Axis Title to "Count".
8. Add an X-axis label. In the PivotChart Tools menu, select **Design > Chart Layouts > Add Chart Element > Axis Titles > Primary Horizontal.** Change the Axis Title to "Date".
9. Rename the Sum Values, as directed in Analysis Tasks 2 and 3.
10. Add a chart title. In the PivotChart Tools menu, select **Design > Chart Layouts > Add Chart Element > Chart Title > Above Chart.** Change the Title to "Number of Sales Orders and Cases Sold by Date".

Analysis Task 5: Add Slicers.

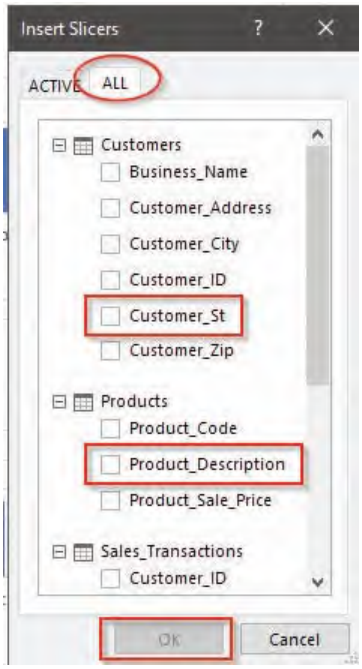
1. Click in any of the pivot tables.
2. Select **Analyze > Filter > Insert Slicer,** as shown in Lab Exhibit 6.4.7.

Lab Exhibit 6.4.7

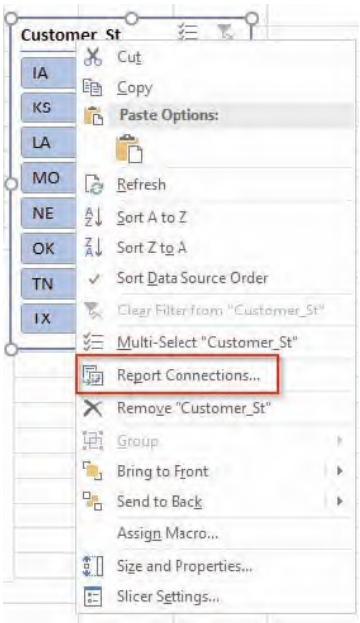


3. In the Insert Slicers dialog box, place check marks next to Customer_St and Product_Description. To see the Customer_St option, you may need to select All, as shown in Lab Exhibit 6.4.8. Click OK.
4. Arrange the Slicers so they are visible.
5. Rearrange the pivot charts if necessary to make a user-friendly dashboard.
6. Right-click anywhere in the Customer_St Slicer. From the options provided, click Report Connections. . ., as shown in Lab Exhibit 6.4.9.
7. Ensure that there is a check mark next to each of the three pivot tables. Click OK. This makes the slicer work on all three pivot charts.
8. Repeat the same action for the Product_Description slicer, but for this one, ensure that there is only a check mark next to PivotTable3 for Count of Sales_Order_ID and Sum of Sales_Order_Quantity_Sold.

Lab Exhibit 6.4.8



Lab Exhibit 6.4.9



9. Rename the Slicers. Right-click the Customer_St filter and select Slicer Settings. . . Rename the slicer "Dashboard-Level Filter: State" in the Caption field; then click OK, as shown in Lab Exhibit 6.4.10.
10. Rename the Product_Description slicer as "Report-Level Filter: Product".

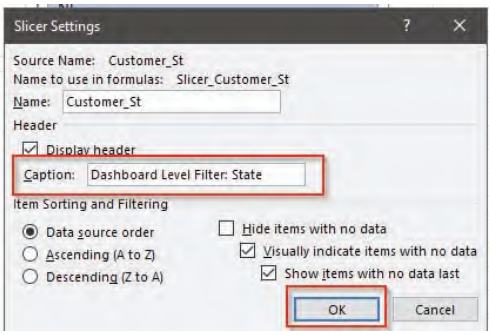


DATA VISUALIZATION

Report the Results: Practice Using the Slicers with the Dashboard

Dashboards serve the purpose of sharing a story by providing information in an easy-to-use and professional format for decision-making. Understanding how the slicers work with the dashboard is important so that you can explain/show the functionality to other users. So, let's practice using slicers to work with the dashboard.

Lab Exhibit 6.4.10



1. Click IA (for Iowa) in the Dashboard-Level Filter: State slicer to see how each pivot table and pivot chart reacts.
2. Clear the filter by clicking the funnel with an X at the top-right corner of the slicer. The pivot charts should go back to the original content.
3. Now click Imperial Stout in the Report-Level Filter: Product slicer to see how only PivotTable3 and PivotChart3 react. PivotTable3 is for Count of Sales_Order_ID and Sum of Sales_Order_Quantity_Sold.

Assessment

1. Use the filters so that your dashboard shows the totals for Imperial Stout sales to Iowa (IA) customers. Take a screenshot of your dashboard, copy it into a Word document named "Lab 6.4 Excel Submission.docx", and label the screenshot Submission 1.
2. Clear all of the filters in your slicers. Take a screenshot of your dashboard, copy it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 6.4 Excel Submission.docx via Connect, if assigned.

Alternate Lab 6.4 Excel

Descriptive Analytics: Creating a Dashboard Using Pivot Tables and Slicers

Apply the same steps in Lab 6.4 to the Alt Lab 6.4 Data.xlsx data set. While Lab 6.4 contains data for transactions with customers, the Alt Lab 6.4 Data contain purchasing transactions with vendors/suppliers for raw materials, which will be converted into the beer that Sláinte will sell to customers. A new data dictionary is presented in the Own the Data Section.

Required

1. Create three pivot tables and pivot charts.
   a. Pivot table 1 and (clustered) column chart: Raw_Materials_Description for Rows and Sum of Quantity_Purchased for Values
   b. Pivot table 2 and (clustered) column chart: Raw_Materials_Description for Rows and Count of Purchase_Order_ID for Values (Count of Purchase_Order_ID indicates how many purchase transactions took place)
   c. Pivot table 3 and line chart: Purchase_Order_Date for Rows and both measures (Sum of Quantity_Purchased and Count of Purchase_Order_ID) for Values
2. Create slicers and modify them so that the first is dashboard-level and the second is report-level. Name the slicers as indicated.
   a. Slicer 1: Dashboard-level Supplier_St (for all pivot tables)
   b. Slicer 2: Report-level Raw_Materials_Description (associated with Pivot Table 3 only)

3. Work with the slicers to present different versions of the reports in the dashboard.
   a. Filter for only Arkansas (AR)
   b. Remove the filter
   c. Filter for only Malt-Wheat

Specify the Questions

1. What quantity of raw materials was purchased from vendors/suppliers, and how many purchase transactions took place in each state?
2. What quantity of raw materials was purchased from vendors/suppliers, and how many purchase transactions took place in each state on each transaction date?

Obtain the Data

This lab uses the data in Alt Lab 6.4 Data.xlsx, which contains vendor transaction data for Sláinte from January through June of 2026.[17] The data dictionary for this file appears below. The data include three worksheets/tables: Purchase_Orders, Suppliers, and Raw_Materials. In the Purchase_Orders worksheet/table, the data include an observation/row for each purchase transaction and type of raw material purchased. So, if a transaction included the purchase of two different raw materials, the data set will have one unique Purchase_Order_ID with two rows (that is, one for each raw material type).

Data Dictionary

Purchase_Orders Table:

Purchase_Order_ID:	Unique purchase transaction identifier
Purchase_Order_Date:	Date of purchase order
Purchase_Employee_ID:	Unique identification number of the employee making the purchase
Supplier_ID:	Unique supplier/vendor identifier
Raw_Materials_Code:	Unique raw material product identifier
Quantity_Purchased:	Pounds of raw material purchased in the transaction

Suppliers Table:

Supplier_ID:	Unique supplier/vendor identifier
Supplier_Company_Name:	Supplier/vendor name
Supplier_Address:	Supplier/vendor street address
Supplier_City:	Supplier/vendor city
Supplier_St:	Supplier/vendor state
Supplier_Zip:	Supplier/vendor zip code

Raw_Materials Table:

Raw_Materials_Code:	Unique raw material product identifier
Raw_Materials_Description:	Description of raw material
Raw_Materials_Price:	Price of one unit of raw materials

Assessment

1. Use your filters so that your dashboard shows the totals for Malt-Wheat purchases from Arkansas (AR) vendors. Take a screenshot of your dashboard, copy it into a Word document named "Alt Lab 6.4 Excel Submission.docx", and label the screenshot Submission 1.
2. Clear all of the filters in your slicers. Take a screenshot of your dashboard, copy it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 6.4 Excel Submission.docx via Connect, if assigned.

[17]These data are a slightly altered version of Lab 4-5 data in Richardson, Terrell, and Teeter, 2021, *Introduction to Data Analytics for Accounting.* New York, NY: McGraw Hill.

Lab 6.5 Tableau

Descriptive Analytics: Creating a Dashboard



DATA VISUALIZATION

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Tableau, Inner Join, Bar Chart, Line Chart, Dashboard

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" Two descriptive techniques used to understand the data are bar charts and line charts. We have worked with Tableau in previous labs, but we have always worked with only one sheet at a time. Now, we will create several Tableau sheets and combine them to create a dashboard, which is an effective way to present a collection of useful visualizations, measures, and tables that allow a company to monitor progress over a period of time or at a particular point in time. An interactive dashboard allows users to dig deep into the data by filtering one or more of visualizations in the dashboard.

Sláinte is a fictional brewery that sells six different beers. It has recently expanded the distribution of its products from one state to nine states, and now its business has begun stabilizing after the expansion. With that stability comes a need for better analysis.

Assume Sláinte has hired you to help management better understand the company's sales data and provide input for its strategic decisions. Management wants all of the analysis placed in an interactive dashboard.

Required

1. Create a dashboard that consists of four Tableau visualizations:
   a. A bar chart for the total number of cases sold by beer type
   b. A bar chart for total number of sales order transactions by beer type
   c. A line chart showing the total number of cases sold and total number of sales orders
   d. A geographic filled map showing number of cases sold in each state
2. Add the following interactive dashboard filters:
   a. Use the first chart, sum of products sold, as a dashboard-level filter.
   b. Use the last chart, the geographic filled map, as a dashboard-level filter.

Specify the Questions

1. How many total cases of each beer were sold?
2. How many sales orders transactions did each beer have?
3. How many cases of beer were sold each day?
4. How many sales order transactions took place each day?
5. How beer cases were sold in each state?

Obtain the Data

This lab uses the data in Lab 6.5 Data.xlsx, which contains sales transaction data for Sláinte from January through June of 2026.[18] The data dictionary for this file appears on the following page. The data include three worksheets/tables: Sales_Transactions, Customers, and Products. In the Sales_Transactions worksheet/table, the data include an observation/row for each sales transaction and product sold. So, if a transaction included the sale of two products, the data set will have one unique Sales_Order_ID with two rows (that is, one for each product).

[18]These data are a slightly altered version of Lab 3-4 data in Richardson, Terrell, and Teeter. 2021. *Introduction to Data Analytics for Accounting.* New York, NY: McGraw Hill.

Data Dictionary

Sales_Transactions Table:

Sales_Order_ID:	Unique sales transaction identifier
Sales_Order_Date:	Date of sales order
Sales_Employee_ID:	Unique identification number of the employee making the sale
Customer_ID:	Unique customer identifier
Product_Code:	Unique product identifier
Sales_Order_Quantity_Sold:	Number of cases sold in the sales transaction

Customers Table:

Customer_ID:	Unique customer identifier
Business Name:	Customer name
Customer_Address:	Customer street address
Customer_City:	Customer city
Customer_St:	Customer state
Customer_Zip:	Customer zip code

Products Table:

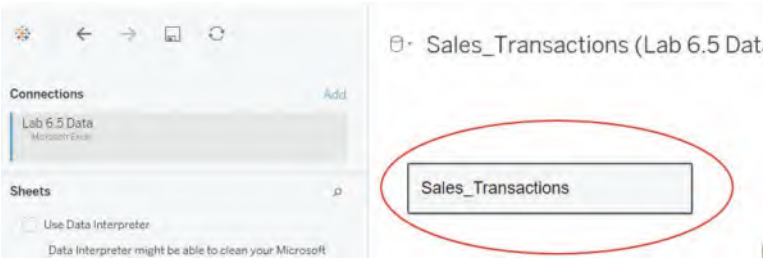
Product_Code:	Unique product identifier
Product_Description:	Type of beer
Product_Sale_Price:	Price of a case of beer

Obtain Task 1: Load the Data into Tableau and Prepare the Data for Use.

1. Open Tableau.
2. Under Connect, select Microsoft Excel.
3. In the Open window, browse to the Lab 6.5 Data.xlsx file.
4. Click Open.
5. In Tableau 2020.2 and higher, the data source pane defaults to a logical layer for building relationships. For this lab, we need to switch the view to the physical layer to work with joins. To do so, add the Sales_Transactions sheet to the logical layer by following these directions: Drag Sales_Transactions to the Data Source pane, then double-click the Sales_Transactions label in the logical layer (as shown in Lab Exhibit 6.5T.1).
6. Drag all the remaining sheets/tables to the Data Connection page. Ensure that you have inner joins created for Sales_Transactions Customers and for Sales_Transactions Products, as shown in Lab Exhibit 6.5T.2.
7. If Tableau imported the Customer_St field as text, change the data type to **Geographic Role > State/Province**, as shown in Lab Exhibit 6.5T.3. Change Sales Order ID and Sales Employee ID to text. Your data should be ready to use.

Lab Exhibit 6.5T.1

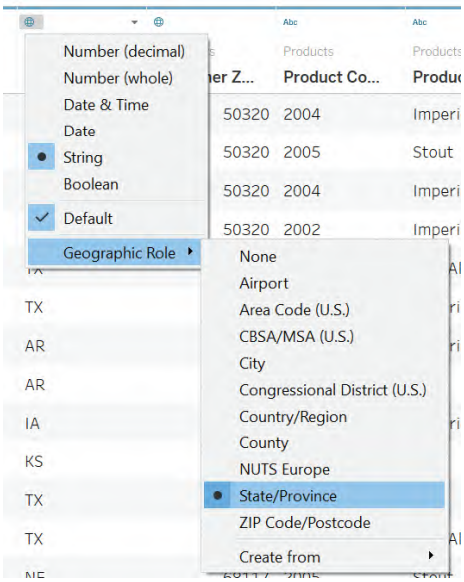
Tableau Software





Lab Exhibit 6.5T.2

Tableau Software



Lab Exhibit 6.5T.3

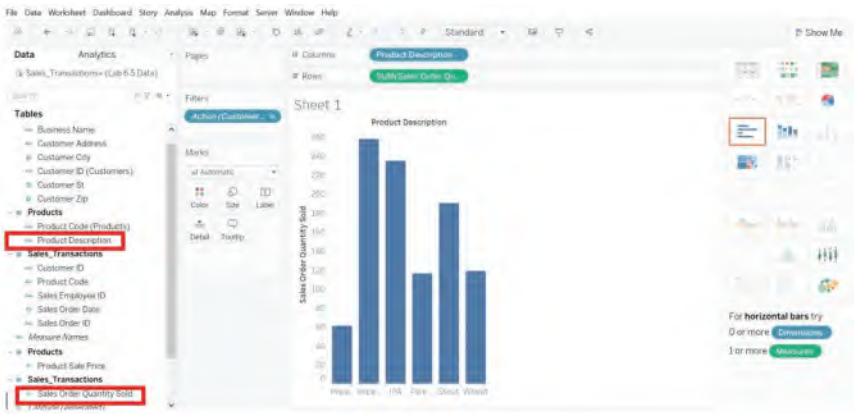
Tableau Software

Analyze the Data

Analysis Task 1: Create the First of Four Visualizations: Number of Cases Sold by Beer Type.

We will create four separate data visualizations, or sheets, that we will combine into one dashboard. Analysis Task 1 creates the first visualization for Sum of Products Sold.

1. Click on Sheet 1.
2. Drag Sales Order Quantity Sold from the Data tab to the Rows shelf.
3. Drag Product Description from the Data tab to the Columns shelf.
4. Tableau should default to a column chart, but if it doesn't, use the Show Me tab to create a bar chart (as shown in Lab Exhibit 6.5T.4). Either vertical or horizontal bars are acceptable for this analysis.



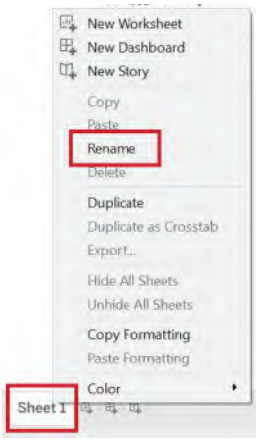
Lab Exhibit 6.5T.4

Tableau Software

5. Right-click on Sheet 1, select Rename as shown in Lab Exhibit 6.5T.5, and type "Number of Cases Sold".

Lab Exhibit 6.5T.5

Tableau Software

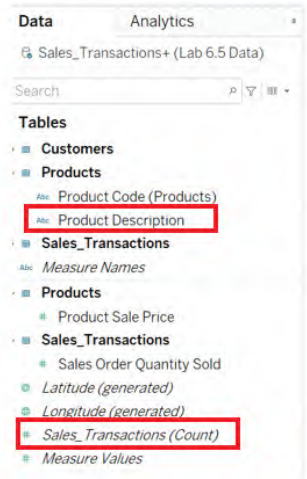


Analysis Task 2: Create the Second of Four Visualizations: Number of Sales Transactions by Beer.

1. Click on New Worksheet to create a new visualization.
2. Double-click on the measure Sales_Transactions (Count).
3. Double-click on Product Description (as shown in Lab Exhibit 6.5T.6) to create a bar chart.

Lab Exhibit 6.5T.6

Tableau Software

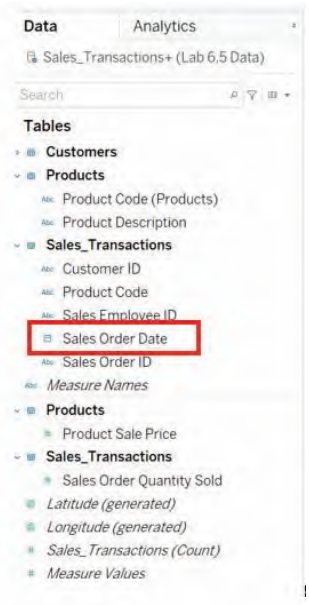


4. Rename this sheet "Number of Sales Order Transactions".

Analysis Task 3: Create the Third of Four Visualizations: Sales by Day.

Let's create a dual line chart for this analysis. We are going to place the Count of Sales Orders by Product over the Sum of Quantity Sold for each day in a dual line chart. In this instance, the line charts provide trend analysis, so the dual axes can be easily interpreted to see the trends of both measures.

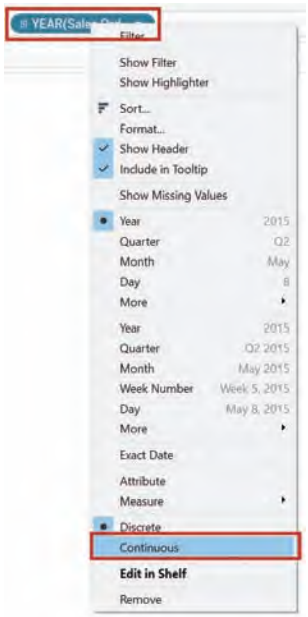
1. Click on New Worksheet.
2. Double-click on Sales Order Date from the Dimensions, as shown in Lab Exhibit 6.5T.7.



Lab Exhibit 6.5T.7

Tableau Software

3. Tableau defaults to interpreting dates as discrete measures. To create a meaningful line chart, transform this variable into a continuous measure. Click the Sales Order Date pill from the Columns shelf and select Continuous, as shown in Lab Exhibit 6.5T.8.



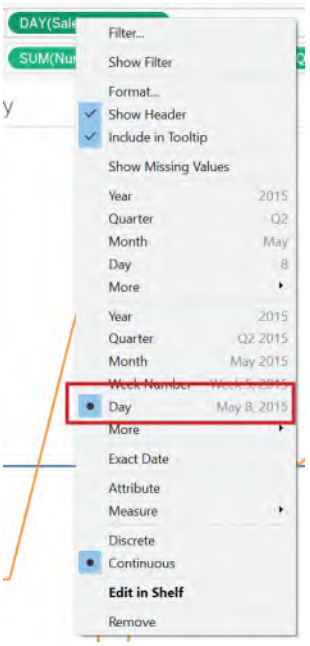
Lab Exhibit 6.5T.8

Tableau Software

4. Click the same Sales Order Date pill again to remove the date hierarchy; select Day, as shown in Lab Exhibit 6.5T.9.
5. To add the measures, double-click on both Sales Order Quantity Sold and Sales_ Transactions (Count), as shown in Lab Exhibit 6.5T.10.

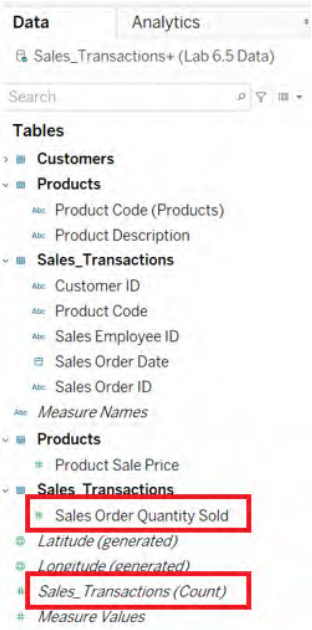
Lab Exhibit 6.5T.9

Tableau Software



Lab Exhibit 6.5T.10

Tableau Software



6. Tableau likely defaulted to showing two separate line charts. To show a dual line chart, click the icon for Dual Line chart from the Show Me tab, as shown in Lab Exhibit 6.5T.11. If the Show Me tab is not showing, click on the Show Me tab to bring it to the front.
7. Rename this sheet "Sales by Day".
8. Label the X-axis "Day". Label the Y-axis "Count of Sales Transactions".
9. The two Y-axes are not the same. To make them the same (or synchronize them), right-click on the Sales Order Quantity Sold axis and check Synchronize Axis.

Analysis Task 4: Create the Fourth of Four Visualizations: Number of Cases Sold by State.

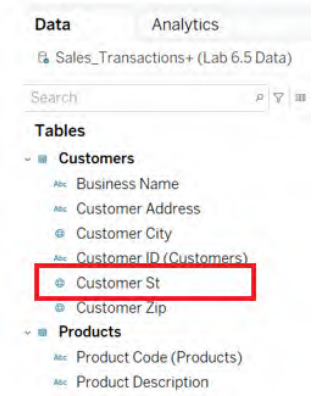
The last data visualization is for geographic data.

1. Click on New Worksheet to open a new data visualization.
2. Double-click on Customer_St from the dimensions, as shown in Lab Exhibit 6.5T.12.



Lab Exhibit 6.5T.11

Tableau Software



Lab Exhibit 6.5T.12

Tableau Software

3. If you adjusted your data type correctly at the beginning of this lab, Tableau should default to a symbol map.
4. We can make this map more interesting by adding a measure. Double-click on Sales Order Quantity Sold from the Measures.
5. To make the map easier to read, select the Filled Map icon from the Show Me tab, as shown in Lab Exhibit 6.5T.13.
6. Rename this sheet "Number of Cases Sold by State".

Analysis Task 5: Create the Dashboard.

1. At the bottom, instead of creating a new sheet, create a new dashboard. Select the icon to the right of the new sheet icon, as shown in Lab Exhibit 6.5T.14.

Lab Exhibit 6.5T.13

Tableau Software



Lab Exhibit 6.5T.14

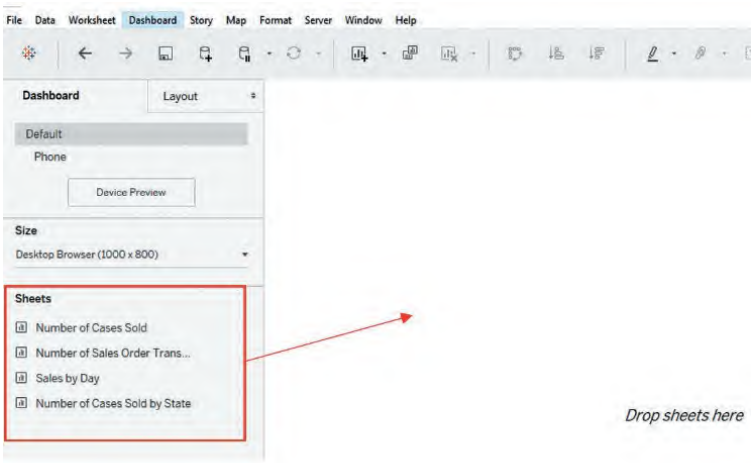
Tableau Software



2. From the Dashboard screen, drag and drop the four sheets to the area that says Drop sheets here, as shown in Lab Exhibit 6.5T.15.

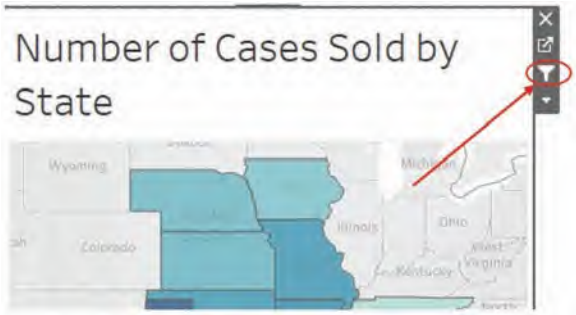
Lab Exhibit 6.5T.15

Tableau Software



Analysis Task 6: Add Dashboard-Level Filters.

1. Once your sheets are arranged in the dashboard, you can click the sheet that you would like to use as a filter. Let's use Number of Cases Sold by State as a filter. Click on the Number of Cases Sold by State visualization.
2. Click the filter button that appears in the top right, as shown in Lab Exhibit 6.5T.16.



Lab Exhibit 6.5T.16

Tableau Software

3. Repeat the same steps for "Number of Cases Sold" sheet. Now both of those sheets act as filters for the dashboard.

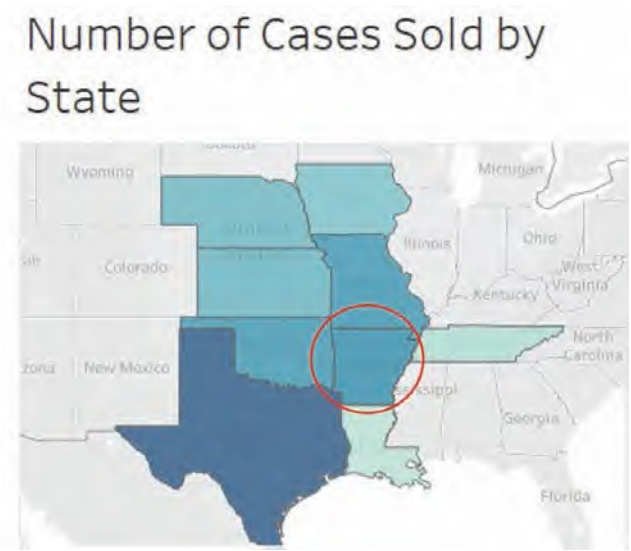
Report the Results: Create Different Views of the Dashboard Data



DATA VISUALIZATION

You have now created an interactive dashboard to analyze sales and product questions with this data set. We can use the filters to present different views of our data.

1. Click the state of Arkansas to see how the rest of the dashboard responds to this interactive filter, as shown in Lab Exhibit 6.5T.17.



Lab Exhibit 6.5T.17

Tableau Software

2. To remove the filter, click anywhere outside of the states with data in them.
3. To select more than one filter, use the ctrl key. For example, click on the state of Texas. Hold the ctrl key and click on the state of Oklahoma. The dashboard should show the combined results for Texas and Oklahoma now. To clear the filter, click anywhere outside of the states with data in them.
4. Practice using the Number of Cases Sold chart as a filter. Click on the bar for Imperial IPA.

Assessment

1. Take a screenshot of your dashboard with the Arkansas (AR) filter on, copy it to a Word document named Lab 6.5 Tableau Submission.docx, and label the screenshot Submission 1.
2. Clear all of the filters in your dashboard. Take a screenshot of your dashboard, copy it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 6.5 Tableau Submission.docx file via Connect, if assigned.

Lab 6.5 Power BI



DATA VISUALIZATION

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Descriptive Analytics: Creating a Dashboard

Keywords

Power BI, Inner Join, Bar Chart, Line Chart, Dashboard

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" Two descriptive techniques used to understand the data are bar charts and line charts. We have worked with Power BI in previous labs, but we have always worked with one core visualization. Now, we will create several Power BI visualizations and combine them to create a dashboard, which is an effective way to present a collection of useful visualizations, measures, and tables that allow a company to monitor progress over a period or time or at a particular point in time. An interactive dashboard allows users to dig deep into the data by filtering one or more of visualizations in the dashboard.

Sláinte is a fictional brewery that sells six different beers. It has recently expanded the distribution of its products from one state to nine states, and now its business has begun stabilizing after the expansion. With that stability comes a need for better analysis.

Assume Sláinte has hired you to help management better understand the company's sales data and provide input for its strategic decisions. Management wants all of the analysis placed in an interactive dashboard.

Required

1. Create a dashboard that consists of four Power BI visualizations:
   a. A bar chart for the total number of cases sold by beer type
   b. A bar chart for total number of sales order transactions by beer type
   c. A line chart showing the total number of cases sold and total number of sales orders
   d. A geographic filled map showing number of cases sold in each state
2. Add the following interactive dashboard filters:
   a. Use the first chart, sum of products sold, as a dashboard-level filter.
   b. Use the last chart, the geographic filled map, as a dashboard-level filter.

Specify the Questions

1. How many total cases of each beer were sold?
2. How many sales orders transactions did each beer have?
3. How many cases of beer were sold each day?
4. How many sales order transactions took place each day?
5. How beer cases were sold in each state?

Obtain the Data

This lab uses the data in Lab 6.5 Data.xlsx, which contains sales transaction data for Sláinte from January through June of 2026.[19] The data dictionary for this file appears below. The data include three worksheets/tables: Sales_Transactions, Customers, and Products. In the Sales_Transaction worksheet/table, the data include an observation/row for each sales transaction and product sold. So, if a transaction included the sale of two products, the data set will have one unique Sales_Order_ID with two rows (that is, one for each product).

[19]These data are a slightly altered version of Lab 3-4 data in Richardson, Terrell, and Teeter. 2021. *Introduction to Data Analytics for Accounting.* New York, NY: McGraw Hill.

Data Dictionary

Sales_Transaction Table:

Sales_Order_ID:	Unique sales transaction identifier
Sales_Order_Date:	Date of sales order
Sales_Employee_ID:	Identification number of the employee making the sale
Customer_ID:	Unique customer identifier
Product_Code:	Unique product identifier
Sales_Order_Quantity_Sold:	Number of cases sold in the sales transaction

Customers Table:

Customer_ID:	Unique customer identifier
Business Name:	Customer name
Customer_Address:	Customer street address
Customer_City:	Customer city
Customer_St:	Customer state
Customer_Zip:	Customer zip code

Products Table:

Product_Code:	Unique product identifier
Product_Description:	Type of beer
Product_Sale_Price:	Price of a case of beer

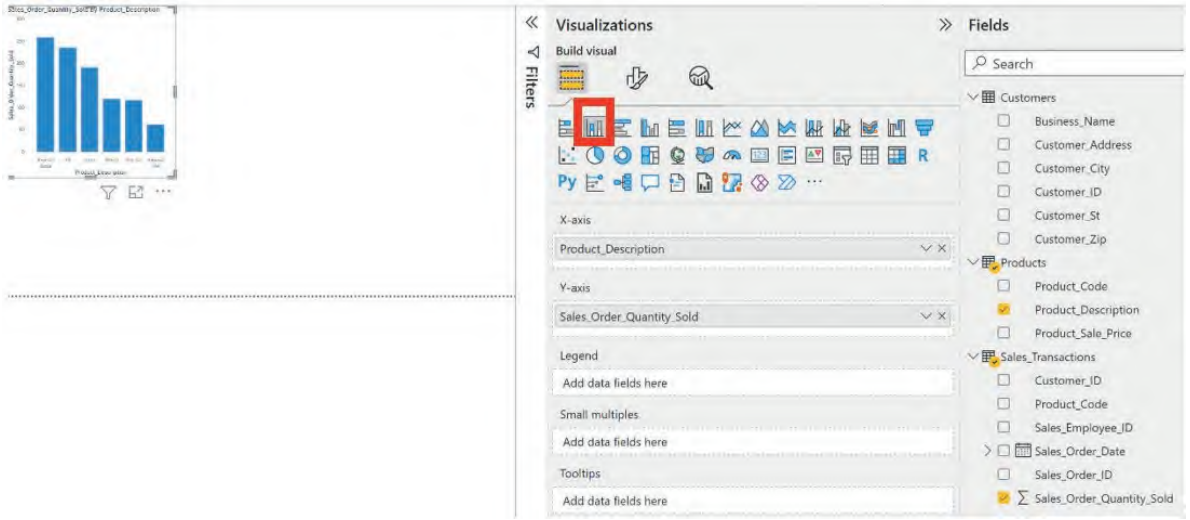
Obtain Task 1: Load the Data into Power BI and Prepare the Data for Use.

1. Open Power BI.
2. Click Get Data.
3. Select Connect.
4. In the Open window, browse to the Lab 6.5 Data.xlsx file.
5. Click Open.
6. Place check marks next to each table name (Customers, Products, and Sales_Transactions) and click Load.
7. Power BI loaded Customer_ID, Product_Code, Sales_Order_ID, and Sales_Employee_ID as Whole numbers. In the Fields pane, highlight Customer_ID in the Customers table. The Column tools tab should appear. Change the Data type to text. Repeat for all instances of the four variables in the Customers, Products, and Sales_Transactions tables.
8. To determine the number of sales orders, create a new column as follows: Click the Sales_Transactions table from the Fields pane, click the Table tools tab on the ribbon, and select New column. In the field that appears, type the following function: NumofSales = COUNT(Sales_Transactions[Sales_Order_ID]) and click Enter.
9. The new column, NumofSales, should appear in the fields beneath Sales_Transactions in the field list.

Analyze the Data

Analysis Task 1: Create the First of Four Visualizations: Number of Cases Sold by Beer Type.

1. From the Visualizations pane, verify that the Build visual icon is highlighted, and click the Stacked column chart to add it to your page.
2. From the Fields pane, check Sales_Order_Quantity_Sold and Product_Description to add them to the Columns chart.
3. Resize the chart to fill the top-left quarter of the page, as shown in Lab Exhibit 6.5P.1



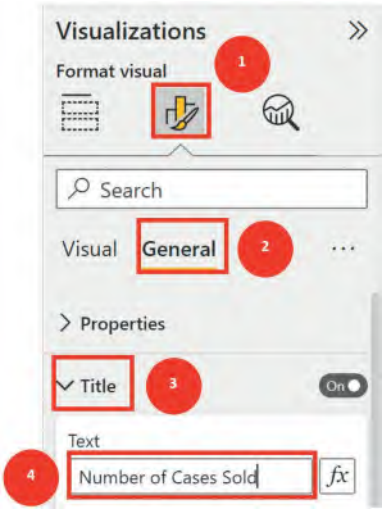
Lab Exhibit 6.5P.1

Microsoft Power BI

4. In the Visualizations pane, click the Format visual icon, click General, expand Title, and type "Number of Cases Sold" in the Title text box, as shown in Lab Exhibit 6.5P.2.

Lab Exhibit 6.5P.2

Microsoft Power BI

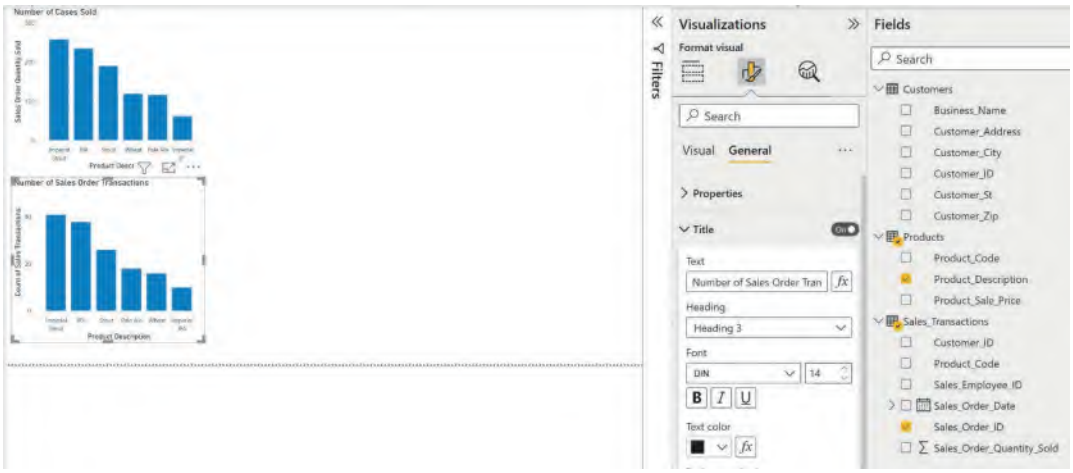


5. Let's change the Y-axis label. Click on the bar chart visualization. Then, in the Visualizations pane, click on the Build visual icon. Expand the box under the Y-axis, select Rename for this visual, and change the default title "Sales_Order_Quantity_Sold" to "Sales Order Quantity Sold". Click Enter.
6. Let's change the X-axis label. Expand the box under the X-axis, select Rename for this visual, and change the default title "Product_Description" to "Product Description". Click Enter.

Analysis Task 2: Create the Second of Four Visualizations: Number of Sales Transactions by Beer.

1. Ensure that the bar chart is not selected. From the Visualizations pane with the Build visual icon selected, add a second stacked column chart to your page.
2. From the Fields pane, drag Product_Description to the X-axis box (Visualizations pane, Build visual) and drag Sales_Order_ID to the Y-axis (Visualizations pane, Build visual). Sales_Order_ID should default to Count.

3. Label the Y-axis as “Count of Sales Transactions” and the X-axis as “Product Description”.
4. Title this chart “Number of Sales Order Transactions”.
5. Resize the chart to fill the bottom-left quarter of the page, as shown in Lab Exhibit 6.5P.3.

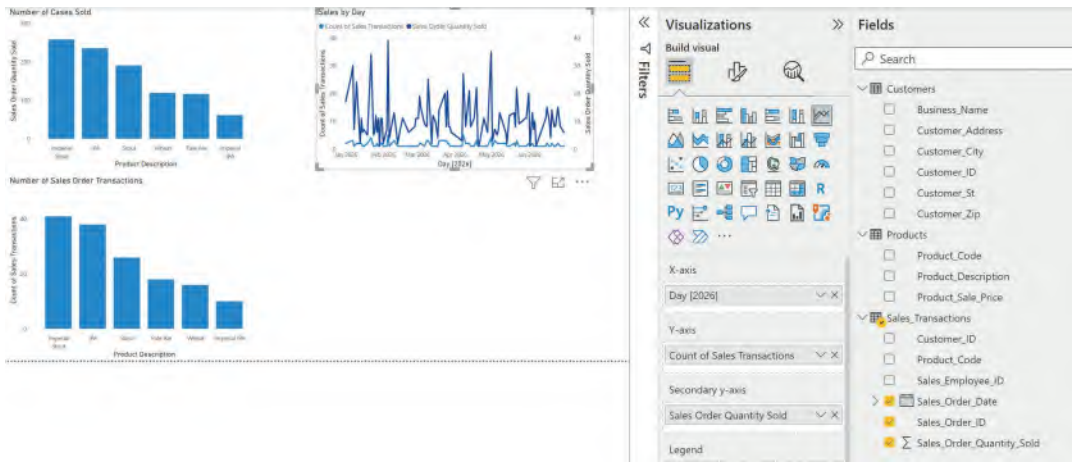


Lab Exhibit 6.5P.3

Microsoft Power BI

Analysis Task 3: Create the Third of Four Visualizations: Sales by Day.

1. Ensure that neither of the bar charts are selected. From the Visualizations pane, click the Build visual icon, and click Line chart to add it to your page.
2. From the Fields pane, drag Sales_Order_Date to the X-axis (in Visualizations, Build visual icon). To show individual days, click the v by Sales_Order_Date and click on Sales_Order_Date to check it (so Date Hierarchy is unchecked). Rename the X-axis (for this visualization) to “Day [2026]”.
3. From the Fields pane, drag Sales_Order_ID to the Y-axis (in Visualizations, Build visual icon). Count of Sales_Order_ID should automatically appear. Rename the Y-axis (for this visualization) to “Count of Sales Transactions”.
4. Let’s add a secondary Y-axis. From the Fields pane, drag Sales_Order_Quantity_Sold to the Secondary y-axis (in Visualizations, Build visual icon). Rename the Secondary y-axis (for this visualization) to “Sales Order Quantity Sold”. To align the Y-axes, click the Format visual icon, expand the Y-axis, set the Minimum to 0, and set the Maximum to 40.
5. Title this chart “Sales by Day”.
6. Resize the chart to fill the top-right quarter of the page, as shown in Lab Exhibit 6.5P.4.

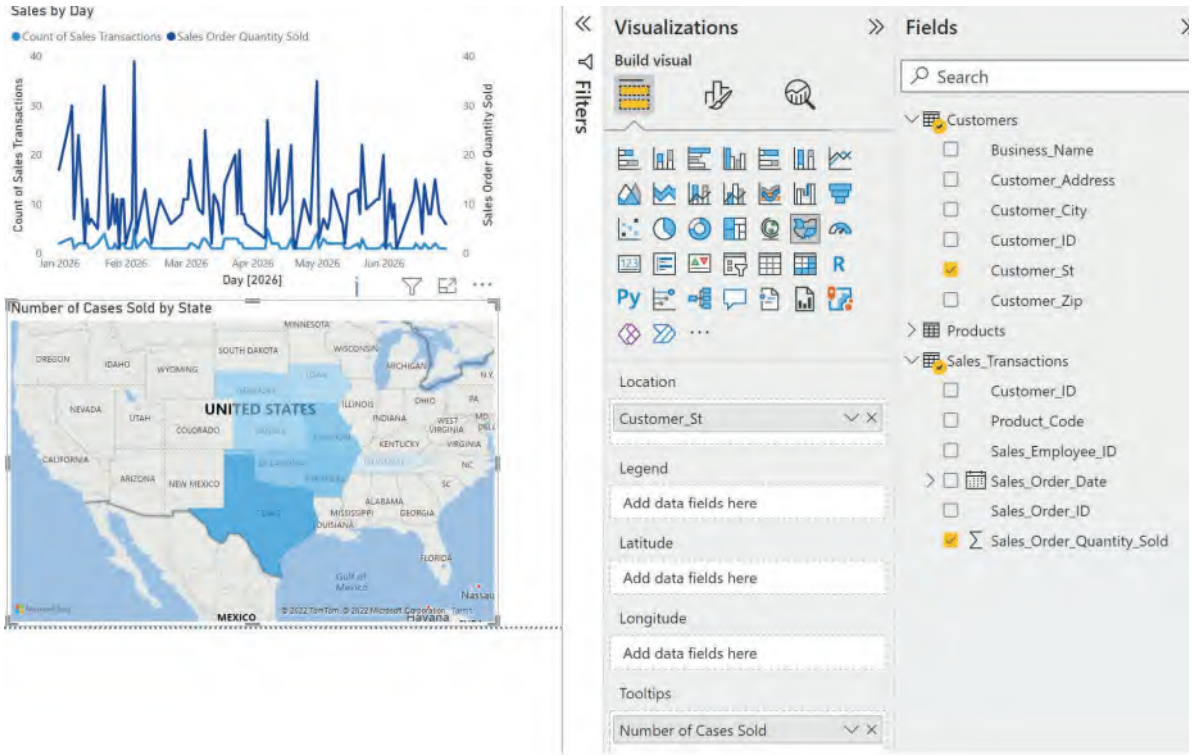


Lab Exhibit 6.5P.4

Microsoft Power BI

Analysis Task 4: Create the Fourth of Four Visualizations: Number of Cases Sold by State.

1. Ensure that none of the visualizations are selected. From the Visualization pane, click the Build visual icon, and then click the Filled map icon to add it to your page.
2. From the Fields pane, drag Customer_St to Location (in the Visualizations pane, Build visual tab.).
3. In the Visualizations pane, click the Format visual icon, expand Fill colors in the Visual tab, and click the Conditional formatting (*fx*) button next to the Default (color). In the Default color - Fill colors dialog box, set What field should we base this on? to Sum of Sales_Order_Quantity_Sold using the drop down arrow (it is in the Sales_Transactions table). Click OK.
4. Title this chart "Number of Cases Sold by State".
5. In the Visualizations pane, click the Build visual icon. Drag Sales_Order_Quantity_Sold to the box under Tooltips. Rename Tooltips to "Number of Cases Sold". This will allow you to see the number of cases sold for each state when you hover the cursor over the state.
6. Resize the chart to fill the bottom-right quarter of the page, as shown in Lab Exhibit 6.5P.5.



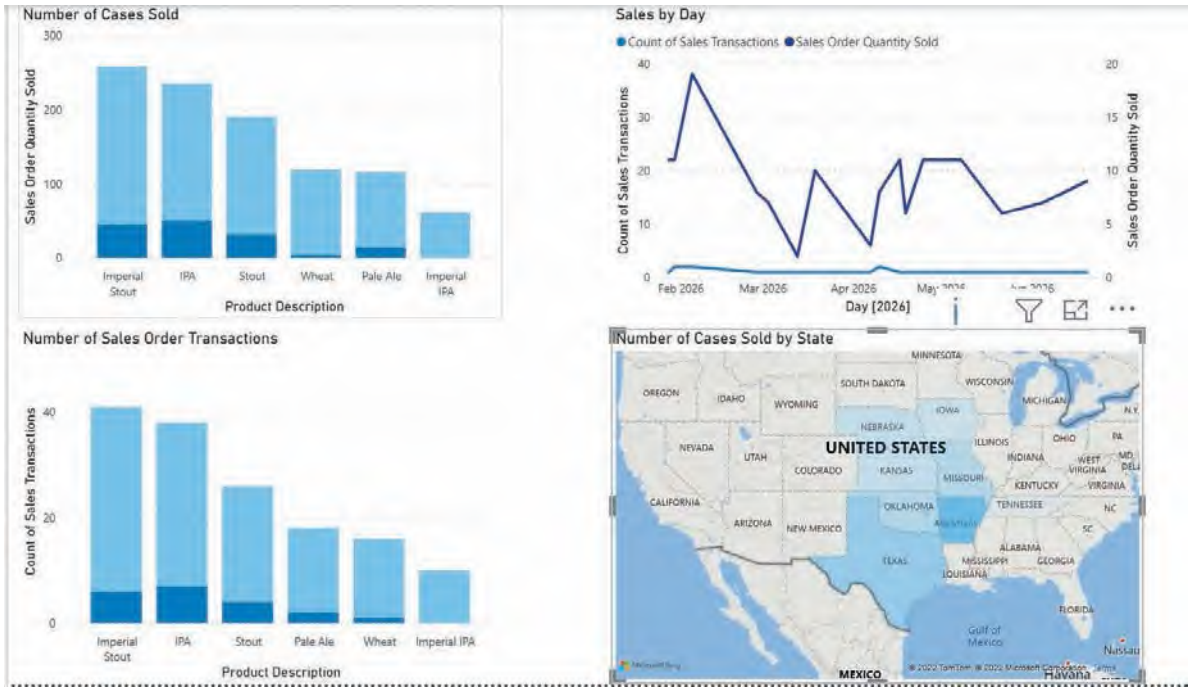
Lab Exhibit 6.5P.5

Microsoft Power BI

Report the Results: Create Different Views of the Dashboard Data

We have now created an interactive dashboard to analyze sales and product questions with this data set. We can use the filters to present different views of our data.

1. Click the state of Arkansas in the Filled map visual to see how the rest of the dashboard responds to this interactive filter, as shown in Lab Exhibit 6.5P.6.
2. To remove the filter, click anywhere outside of the states with data in them.



Lab Exhibit 6.5P.6

Microsoft Power BI

3. To select multiple filters, use the ctrl key. For example, click on the state of Texas. Hold the ctrl key and click on the state of Oklahoma. The page (or dashboard) should show the results for both states now. To remove the filter, click anywhere outside of the states with data in them.
4. Practice using the Number of Cases Sold chart as a filter. Click on the bar for Imperial IPA.
5. Add page-level filters for Number of Cases Sold and State. To add a page-level filter, right-click on the variable in the Fields pane (here, Sales_Order_Quantity_Sold and Customer_St) , and select **Add to filters > Page-level filters**."

Assessment

1. Take a screenshot of the results with the Imperial IPA filter, copy it to a Word document named "Lab 6.5 Power BI Submission.docx", and label the screenshot Submission 1.
2. Clear all the filters. Take a screenshot, copy it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 6.5 Power BI Submission. docx files via Connect, if assigned.

Alternate Lab 6.5

Descriptive Analytics: Creating a Dashboard

Using either Power BI or Tableau, apply the same steps in Lab 6.5 to the Alt Lab 6.5 Data. xlsx data set. While Lab 6.5 contains data for transactions with customers, the Alt Lab 6.5 data set contains data regarding purchasing transactions with vendors/suppliers for raw materials, which will be converted into the beer that Sláinte will sell to customers. A new data dictionary is presented in the Obtain the Data section.

Required

1. For Tableau, create the relationships among the three tables in the data source window. Use Purchase Orders as the initial table.
2. Change the data types if necessary.
3. If you are using Tableau, create the following four sheets and then put them in a dashboard.
   a. Sum of Quantity Purchased by Raw Material (Name: Pounds of Raw Materials Purchased)
   b. Count of Purchase Orders by Raw Material (Name: Number of Purchase Transactions)
   c. Count of Purchase Orders and Sum of Quantity Purchased by Day (Name: Purchases by Day)
   d. Sum of Raw Materials Purchased by State (Name: Pounds of Raw Materials Purchased by State)
4. If you are using Power BI, put the following visualizations on one page.
   a. Sum of Quantity Purchased by Raw Material (Name: Pounds of Raw Materials Purchased)
   b. Count of Purchase Orders by Raw Material (Name: Number of Purchase Transactions)
   c. Count of Purchase Orders and Sum of Quantity Purchased by Day (Name: Purchases by Day)
   d. Sum of Raw Materials Purchased by State (Name: Pounds of Raw Materials Purchased by State) Set "How should we format empty values?" to Don't format.
5. If you are using Power BI, create page-level filters for the following variables:
   a. Raw Materials Description
   b. Supplier State
6. If you are using Tableau, create dashboard-level filters from the following two sheets:
   a. Pounds of Raw Materials Purchased
   b. Pounds of Raw Materials Purchased by State

Specify the Questions

1. How many pounds of each raw material were purchased?
2. How many purchase order transactions did each raw material have?
3. How many pounds of raw material were purchased each day?
4. How many purchase-order transactions took place each day?
5. How many pounds of raw materials were purchased from each state?

Obtain the Data

This lab uses the data in Alt Lab 6.5 Data.xlsx, which contains vendor transaction data for Sláinte from January through June of 2026.[20] The data dictionary for this file appears below. The data include three worksheets/tables: Purchase_Orders, Suppliers, and Raw_Materials. In the Purchase_Orders worksheet/table, the data include an observation/row for each purchase transaction and type of raw material purchased. So, if a transaction included the purchase of two different raw materials, the data set will have one unique Purchase_Order_ID with two rows (that is, one for each raw material type).

[20]These data are a slightly altered version of Lab 4-5 data in Richardson, Terrell, and Teeter, 2021, *Introduction to Data Analytics for Accounting.* New York, NY: McGraw Hill.

Data Dictionary

Purchase_Orders Table:

Purchase_Order_ID:	Unique purchase transaction identifier
Purchase_Order_Date:	Date of purchase order
Purchase_Employee_ID:	Unique identification number of the employee making the purchase
Supplier_ID:	Unique supplier/vendor identifier
Raw_Materials_Code:	Unique raw material product identifier
Quantity_Purchased:	Quantity of raw material purchased in the transaction (pounds)

Suppliers Table:

Supplier_ID:	Unique supplier/vendor identifier
Supplier_Company_Name:	Supplier/vendor name
Supplier_Address:	Supplier/vendor street address
Supplier_City:	Supplier/vendor city
Supplier_St:	Supplier/vendor state
Supplier_Zip:	Supplier/vendor zip code

Raw_Materials Table:

Raw_Materials_Code:	Unique raw material product identifier
Raw_Materials_Description:	Description of raw material
Raw_Materials_Price:	Price of one unit of raw materials

Assessment

1. Add in two filters: a filter for AR (Arkansas) and a filter for Malt Wheat, so that your dashboard shows the totals for Malt Wheat purchases from Arkansas suppliers. Take a screenshot of your dashboard with filters on, copy it to a Word document named "Alt Lab 6.5 Tableau Submission.docx" or "Alt Lab 6.3 Power BI Submission.docx", and label the screenshot Submission 1.
2. Clear all of the filters in your dashboard. Take a screenshot of your dashboard, copy it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 6.5 Submission.docx file via Connect, if assigned.

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Chapter 7

Marketing Analytics

A Look Back

- Chapter 6 discusses why data visualization is important to organizations.
- Several different types of charts can be used in data visualizations.
- Dashboards can be used to present multiple visualizations.

A Look at This Chapter

- Chapters 7 through 10 describe how companies can use business analytics to gain actionable business intelligence (BI) across the four main business functions: marketing, accounting, finance, and operations. These chapters focus on the first four types of analytics: descriptive, diagnostic, predictive, and prescriptive.
- Chapter 7 investigates how business analytics can be used in the marketing function through application of the SOAR model. Marketing helps companies meet customer needs profitably by selecting an appropriate marketing mix, which includes product, price, place, and promotion.

A Look Ahead

- Chapter 8 investigates how business analytics can be used in the accounting function through application of the SOAR model.
- Accounting is the systematic process of measuring, recording, and communicating the financial information that decision-makers need to make decisions.
- Accounting also classifies, verifies, and summarizes business transactions to make the information useful and accessible to decision-makers. A business transaction is the exchange of money for goods and/or services between a company and other parties, including customers and vendors/suppliers.



(Source: B. Watson, 2021)
Marcia Watson/McGraw Hill



(Source: B. Watson, 2022. *The Grotto Serpent.*)
Marcia Watson/McGraw Hill

LEGO: The Building Blocks of a Successful Marketing Campaign

"Everyone has their LEGO story."
—LUCY HANDLEY

Founded in 1934 in a small Danish village, **LEGO** (which means "play well" in Danish) has become the world's most valuable toy company, larger than **Fisher-Price**, **Barbie**, **Nerf**, and **Hasbro**. The company started with one design, a 2 X 4 brick. Today, LEGO manufactures over 3,700 different types of pieces, allowing seemingly unlimited building options. In 2015, LEGO was named the world's most powerful **brand**, which is a phrase, symbol, logo, design, or word that uniquely identifies a company. Creative customers have created both simple and intricate designs, like those shown in the photos here.

How did LEGO become so successful? In addition to having a great product, the company has great marketing. For example, the (online) LEGO Ideas Community allows LEGO fans and builders from around the globe to "imagine, iterate, and evaluate ideas" for new LEGO kits. If your idea receives 10,000 support votes in a single calendar year, the LEGO Review Board will review it. If the Board accepts your design, it will become a real LEGO set—and you will receive a 1% royalty on all sales and licensing revenue.

LEGO Ideas is brilliant marketing built around creating not just a brand but also a *brand community.* It not only creates a continuing buzz about LEGO products, but also generates free product promotion as community members share ideas in social media. In addition, it encourages innovation—a key to the company's mission, which is to inspire and develop the builders of tomorrow. As Jaime Netzer of **Khoros**, a software company that helps other companies develop a brand and social media presence, noted, "This community doesn't just keep brand excitement fresh—it also offers the company a treasure trove of accessible market data." LEGO can use these data for marketing analytics to better understand its current and potential customers, improve LEGO products, and increase sales.

What is *your* LEGO story?

(Sources: Lucy Handley. "How Marketing Built LEGO into the World's Favorite Toy Brand," CNBC Marketing.Media.Money, 2018, https://www.cnbc.com/2018/04/27/LEGO-marketing-strategy-made-it-world-favorite-toy-brand.html. Accessed 09/24/2021. Jaime Netzer. "How LEGO, Sephora, and The American Red Cross are Using Empathy to Drive their Brand Communities," 2018, https://khoros.com/blog/how-LEGO-sephora-and-american-red-cross-are-using-empathy-drive-their-brand. Accessed 09/24/2021.)

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

- **LO 7.1** Explain the role of marketing and the components of the marketing mix (product, price, place, and promotion).
- **LO 7.2** Identify representative marketing questions addressable by business analytics for each component of the marketing mix.
- **LO 7.3** Distinguish between internal and external sources of marketing data.
- **LO 7.4** Describe and provide examples of descriptive analytics used to answer marketing questions.
- **LO 7.5** Describe and provide examples of diagnostic analytics used to answer marketing questions.
- **LO 7.6** Describe and provide examples of predictive analytics used to answer marketing questions.
- **LO 7.7** Describe and provide examples of prescriptive analytics used to answer marketing questions.
- **LO 7.8** Describe how marketing analytics reports results.

LABS FOR CHAPTER 7

LAB 7.1 **Excel:** Descriptive Analytics: Analyzing Company Historical Performance
LAB 7.2 **Excel:** Descriptive Analytics: Using a Pivot Table to Analyze Historical Performance by Product Size and Year
LAB 7.3 **Tableau:** Descriptive Analytics: Using a Histogram to Evaluate Process Time
LAB 7.3 **Power BI:** Descriptive Analytics: Using a Histogram to Evaluate Process Time
LAB 7.4 **Excel:** Diagnostic Analytics: Analyzing the Steps in the Sales Process with a Sales Funnel Chart
LAB 7.5 **Tableau:** Diagnostic Analytics: Examining Pricing Strategy with Cluster Analysis
LAB 7.6 **Excel:** Predictive Analytics: Predicting Sales Revenue from Advertising Expense
LAB 7.6 **Tableau:** Predictive Analytics: Predicting Sales Revenue from Advertising Expense
LAB 7.7 **Excel:** Prescriptive Analytics: Calculating Internet CPM Rate Using Goal Seek
LAB 7.8 **Excel:** Prescriptive Analytics: Calculating

LO 7.1

Explain the role of marketing and the components of the marketing mix (product, price, place, and promotion).

7.1 WHAT IS MARKETING?

"Your most unhappy customers are your greatest source of learning."
—**BILL GATES**, co-founder of **Microsoft**, *Business @ the Speed of Thought* (1999)

Marketing is essential to the success of a business. Without marketing, many companies would fail due to a lack of sales.

Marketing is the "activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large."[1] Marketing encompasses many processes, including (1) developing a product or service, (2) making potential customers aware of the product/service, (3) making customers want the product/service, and (4) selling the product/service to customers.

How do managers use marketing? As noted in the opening story, **LEGO** uses customer ideas to create new products. **Nike** advertises with its iconic slogan "Just Do It," which creates product awareness by inspiring customers to push themselves beyond their limits—while they wear their newly purchased Nike gear. **Carvana** transformed the used-car buying experience by allowing customers to use the internet to shop, finance, and trade in cars. Carvana also delivers cars to a customer's house and allows customers to pick up their newly purchased car from a 24-hour car vending machine.

Technically, throughout this chapter, we should differentiate between consumers and customers. *Consumers* use or consume products but don't necessarily buy them. *Customers* buy products but don't necessary use or consume them. For example, if a restaurant buys coffee beans, the restaurant is the customer but not the consumer. The consumer is the person who buys the cup of coffee from the restaurant. For the sake of simplicity, we use the word *customer* throughout this chapter.

The Marketing Mix[2]

To sell products and services to their targeted customers, companies implement their selected marketing mix. The **marketing mix** refers to the key choices that a company makes to bring a product/service to market. It consists of the "the four Ps": product, price, place, and promotion.

The **product** is the good or service that a company offers to customers. To be successful, a product must fulfill some customer demand by incorporating features that customers think they need. When developing a product, the company needs to plan ahead for the different phases of the **product life cycle,** or the four phases that products go through, beginning with development and ending with decline/market removal. The four phases are introduction, growth, maturity, and decline. Each phase has its own price, promotion, and place strategies. In addition, the product features may change over the life cycle. The type of product affects the other three Ps (price, place, and promotion).

Price is the amount that a customer pays for a product. Setting a price requires looking at the costs to produce and distribute the product, the prices of similar items, and the product's perceived value. If the product is priced higher than its perceived value, customers won't buy it. However, sometimes companies use higher prices to make a product seem like a luxury item, which increases the perceived value to at least some customers. Price can affect not only the place where the product is sold but also how the product is promoted.

Place identifies where a company sells a product (for example, at bricks-and-mortar stores, on a website, at company-owned stores) and how the product is delivered to the market.

[1]Source: C. Shane Hunt, John E. Mello, and George D. Deitz. 2021. *Marketing,* 3rd ed. New York, NY: McGraw Hill Education, p. 4.
[2]Source: C. Shane Hunt, John E. Mello, and George D. Deitz. 2021. *Marketing,* 3rd ed. New York, NY: McGraw Hill Education.

Distribution is a key component of place. Generally, a product should be sold in the place(s) where target customers shop. Place can affect the promotions that are used for a product.

Promotion refers to all of the marketing communication strategies and techniques used to convince customers that they need the product at the specified price. It includes advertising, public relations, sales promotions, special offers, and social media.

More recently, additional Ps have been added to the list:

(1) *Planning* refers to the alignment between corporate strategy and marketing strategy.
(2) *Process* refers to the flow of activities that take place when the business interacts with a customer.
(3) *People* refers to having the right employees in terms of skills sets, leadership qualities, and other characteristics.
(4) *Physical evidence* refers to the combination of the physical environment in which the product is delivered (for example, fast-food restaurant versus a five-star restaurant) and any tangible items that may help performance or communication (for example, pamphlets and user manuals).

In this chapter, we focus on the core Ps: product, price, place, and promotion.

Let's look at how the four Ps work for **LEGO**. LEGO's main *product* is toy bricks (mostly sold in sets now) that are marketed to middle-class and upper-class families. Intricate LEGO toys can be costly to produce, and the company charges a slight premium for its products. However, the company does not set the *price* too high, because it wants middle-class families to be able to buy LEGO sets. With respect to *place,* LEGO has production units in the United Kingdom, United States, Singapore, China, and Czech Republic. LEGO has about 740 retail stores worldwide. Its products are also available in department stores, toy stores, movie theaters, amusement parks, big-box stores, and online at **Amazon** and other internet retailers. With respect to *promotion,* LEGO has kept its brand visible by developing additional products, including amusement parks, movies, books, video and board games, clothes, and interactive online content. All of these promotions are carefully designed to promote the buying of LEGO bricks and LEGO sets.[3]

PROGRESS CHECK

1. What are the four components of the marketing mix? What does each represent?
2. How are the components of the marketing mix related?
3. Identify additional ways LEGO can improve its marketing mix.

7.2 SPECIFYING THE MARKETING QUESTION

LO 7.2

Identify representative marketing questions addressable by business analytics for each component of the marketing mix.

"The goal is to turn data into information, and information into insight."

—**CARLY FIORINA**, Former CEO of **HP**

Exhibit 7.1 provides a few basic examples of detailed marketing questions, according to analytics type (descriptive, diagnostic, predictive, and prescriptive), for the four main components of the marketing mix. Be sure to read the full exhibit to get an overview of the many marketing-based questions that managers ask data analysts to answer.

Once the question is determined, companies can use **marketing analytics**, "the processes and technologies that enable marketers to evaluate the success of their marketing

[3]Sources: mbasKool.com, "Lego Marketing Strategy & Marketing Mix (4Ps)," 2017, https://www.mbaskool.com/marketing-mix/products/17385-LEGO.html. Accessed 09/24/2021; Marketing91, "Marketing Mix Of Lego—Lego Marketing Mix," 2019, https://www.marketing91.com/marketing-mix-LEGO/. Accessed 09/24/2021.

	Marketing Mix Component			
Business Analytics Type	**Product**	**Price**	**Place**	**Promotion**
Descriptive: What happened? What is happening?	How many units were sold in the period? What was the average sales revenue from a product for each month? What are the top 10 products in terms of sales and profit?	What was the average price of the product each month? What is the highest-priced product? What is the average price of similar products in the industry?	How many units were sold in each region? How many units were sold in each store? How many units were sold online? How many webpage views resulted in a sale?	What is the monthly expenditure on promotions for each product line? What were sales after a new commercial aired? Did a coupon offer have a positive ROI (return on investment)?
Diagnostic: Why did it happen? What are the reasons for the past results? Why are the results different from expectations?	Why did sales volume change from last period? Why did sales of a particular product change from last period?	How and why did units sold change when we lowered the price of a product? How did product sales change when a competitor lowered prices?	Why did sales decrease in one region but increase in another region? Does weather affect regional sales?	Why did webpage views change? How did a sales promotion or advertisement affect the purchasing behavior of different types of customers?
Predictive: What will happen in the future? What is the probability something will happen? Can we forecast what will happen?	What products will customers buy together in the future? What will be unit sales in the different phases of the product life cycle?	What are the estimated unit sales and total sales revenue for a price change?	What will unit sales be, based on season and location? What will unit sales be, based on marketing channel?	After brand repositioning, what results should be expected? How will a promotion affect each marketing channel?
Prescriptive: What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints?	What features of the product should be changed to maximize sales revenue? Which products should be discontinued?	What is the optimal price to charge for a product if we want to increase sales volume? What is the optimal price to charge for a product if we want to increase total sales revenue?	What is the optimal product assortment at each store? Which marketing channel should carry/not carry products?	How can we best target promotions? What steps/promotions should be taken to increase repeat purchases? How can we optimize advertisements across marketing channels?

Exhibit 7.1
Examples of Marketing Questions by Business Analytics Type and Marketing Mix Component

initiatives by measuring performance using important business metrics."[4] Marketing analytics often compares the results from different **marketing channels**, which are final product distribution points, such as wholesalers, online stores, or bricks-and-mortar retail stores. Marketing analytics considers the results from all marketing channels over a period of time to give managers a fuller understanding of the impact of marketing efforts to make the right product available to the right persons at the right time in the right place and at the right price.

[4]Source: SAS, "Marketing Analytics: What It Is and Why It Matters," 2021.
Jonathan Moran, "What's In a Name . . . Marketing Analytics," https://blogs.sas.com/content/customeranalytics/2013/03/07/whats-in-a-name-marketing-analytics/, March 7, 2013. Accessed 1/15/2022.

MARKETING ANALYTICS AND YOUR CAREER

One job that makes extensive use of marketing analytics is the market research analyst. These analysts gather data from a variety of sources and use statistical techniques and software to evaluate the effectiveness of marketing campaigns, forecast sales, and suggest pricing. The median salary for a market research analyst is $65,810, but senior-level people may earn more than $100,000. Job growth for market research analysts is projected to be 22 percent from 2020 to 2030, which is much faster than the average for all occupations.[5] Other marketing jobs requiring marketing analytics skills include marketing manager, brand manager, digital marketing specialist, creative director, and director of marketing strategy. Clearly, there are many great opportunities for people with skills in marketing analytics.

PROGRESS CHECK

4. Give an example of a marketing question associated with predictive analytics.
5. Which component of the marketing mix (that is, which one of the four Ps) do you think is the most important for a company to be successful?

7.3 OBTAIN THE DATA: WHAT MARKETING DATA ARE AVAILABLE? A DISCUSSION OF MARKETING DATA SOURCES

LO 7.3

Distinguish between internal and external sources of marketing data.

"Marketing without data is like driving with your eyes closed."

—DAN ZARRELLA, **Hubspot**

A huge amount of data is available to help answer the marketing questions in Exhibit 7.1 and many other marketing-based questions, too. This section explains the two key sources of marketing data: internal data and external data. Companies can use a variety of software tools. including R, Python, and SQL, to obtain and prepare these data for analysis.

Internal Marketing Data

Companies have vast amounts of internally stored data that can be used for marketing analytics. Chapter 2 described the most common source of a company's internal data: the enterprise resource planning (ERP) system or enterprise system (ES). Some companies use a large ERP/ES, such as **SAP** or **Oracle**. Smaller companies tend to use **QuickBooks** or **Peachtree**.

The ERP/ES stores all sales transaction data and pricing data. Transaction data include the customer, the amount, the items/services sold, the payment type and amount, the date of the transaction, and the cost of the product sold. Thus, the ERP/ES contains total sales and total cost information for a particular product, as well as the total dollars spent on advertising, promotions, and other marketing campaigns.

[5]Source: U.S. Bureau of Labor Statistics, "Market Research Analysts," *Occupational Outlook Handbook,* September 8, 2021, https://www.bls.gov/ooh/business-and-financial/market-research-analysts.htm. Accessed 9/26/2021.

Some companies also use a **customer relationship management (CRM) system**, which can be part of the ERP/ES or a standalone system. Popular CRM systems are available from **Hubspot**, **Mailchimp**, **Oracle**, **SAP**, and **Salesforce**. A CRM system manages the company's interactions with past, current, and potential customers. It stores contact information, records sales and service issues, helps manage marketing campaigns to identify potential sales opportunities, and manages/records communications with customers across marketing channels. By storing customer data in one shareable location, it helps companies increase sales by identifying cross-selling and upselling opportunities, improving service experiences by tailoring them to customer preferences, and personalizing promotions. For example, a CRM system can assist the company in sending personalized emails to customers, track if the emails were opened, track customer actions such as link clicks and purchases, and even predict what time of day a specific customer is more likely to open an email.

Customer loyalty programs are one way to augment the information collected by the CRM. Customers like customer loyalty programs because they typically come with perks, such as discounts or free shipping. From the company's perspective, customer loyalty programs create a treasure trove of information. They allow the tracking of individual customer buying patterns, tracking the time and day of each purchase, how often a product is purchased, how often a customer buys certain products together, and the customer's response to different coupons and marketing campaigns.

SOARing to Success

Teaching Companies to Offer You a Discount

In the past, one of the authors bought from **Home Depot** only when she or her husband had a coupon for 10 percent or more off. They generally received a coupon every couple of months. Now, they both receive coupons almost every week via email and traditional "snail mail." Home Depot's CRM system must have determined that they purchase only when they have a coupon in hand. So, now Home Depot sends them coupons galore! By the way, the strategy worked. They go to Home Depot much more frequently now to purchase a wider variety of items. What other marketing promotions could Home Depot use to increase sales? Which other companies use a similar strategy?

Company websites also provide a wealth of information, especially when combined with **web analytics**, which extracts and categorizes website quantitative and qualitative data to identify and analyze patterns and trends. Web analytics analyzes not only website traffic (the number of visits a company's website gets over a specific period time), but also page views related to products, links clicked, return visits, and user/customer comments. Web analytics also tracks the results of *A/B testing,* in which two different company websites are presented to two different groups of customers to determine which version has the higher **conversion rate**, or the highest percentage of website users who actually purchase from the website.

Web analytics can also monitor potential traffic sources, which are other websites that led a potential customer to the company's website. Traffic sources include search engines such as **Google**, other websites, email campaigns, and links from social media. The information provided can help a company adjust its marketing strategy. For example, if search engine traffic is low, then the company needs to engage in **search engine optimization (SEO)** by selecting popular keywords, terms, and phrases that direct search engines to the company's website.

MINI CASE: Using Google Analytics to Get Started with Web Analytics

Google Analytics, a free service provided by Google that tells a company where its website traffic originates, is a good place to start working with web analytics. Exhibit 7.2 shows one of the many reports that Google Analytics can generate. The abbreviation "cpc" in the first row under the "Source/Medium" heading stands for cost per click, indicating that the user clicked on a company-paid ad on another website designed to drive traffic to the company website. The word "organic" indicates unpaid/free listings on search engines. For example, typing "chocolate chip cookies delivery" into **Google**'s search bar returns **Tiff's Treats Cookie Delivery** website. The word "referral" indicates traffic from a variety of other sources, including Google Group posts and other pages on Google-related websites.

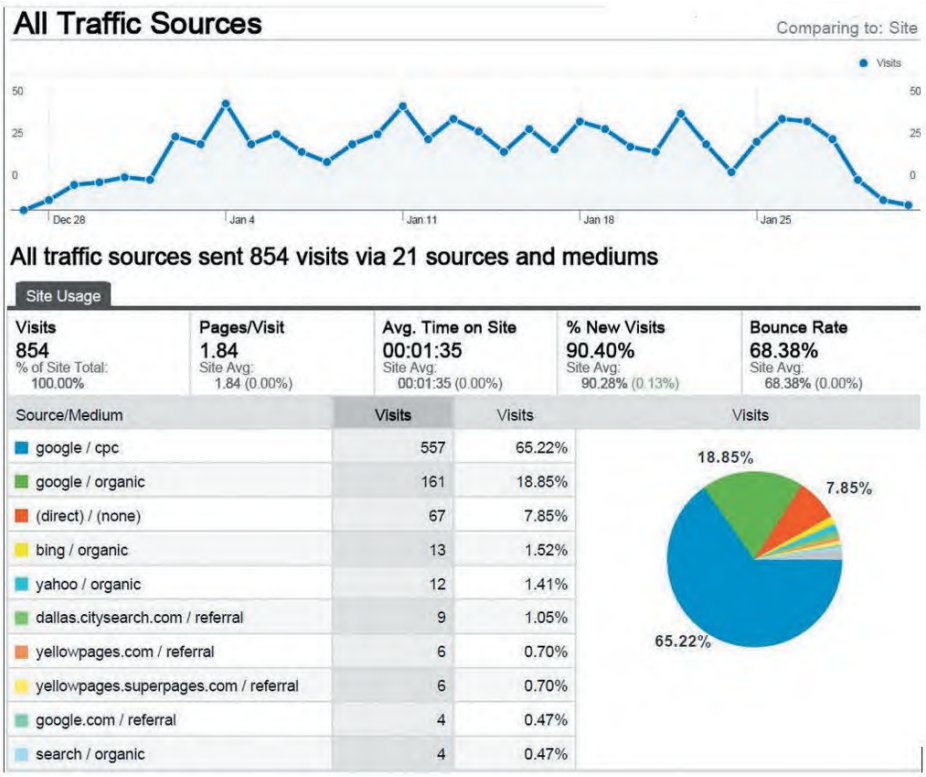


Exhibit 7.2
Example of Google Analytics

(Source: Molly Clubb, "Success Stories from Spring Sales: Keeping Your Momentum Going," 2014, https://www.hired-handsoftware.com/resources/blog/hired-hand/2014/05/27/success-stories-from-spring-sales-keeping-your-momentum-going. Accessed 09/24/2021.)

Decision-Making with Data

Use Exhibit 7.2 to answer the following questions. Where does most of the company's web traffic originate during the period analyzed? Is the company spending its advertising dollars wisely? Should the company be concerned about recent web traffic? Why or why not?

External Marketing Data

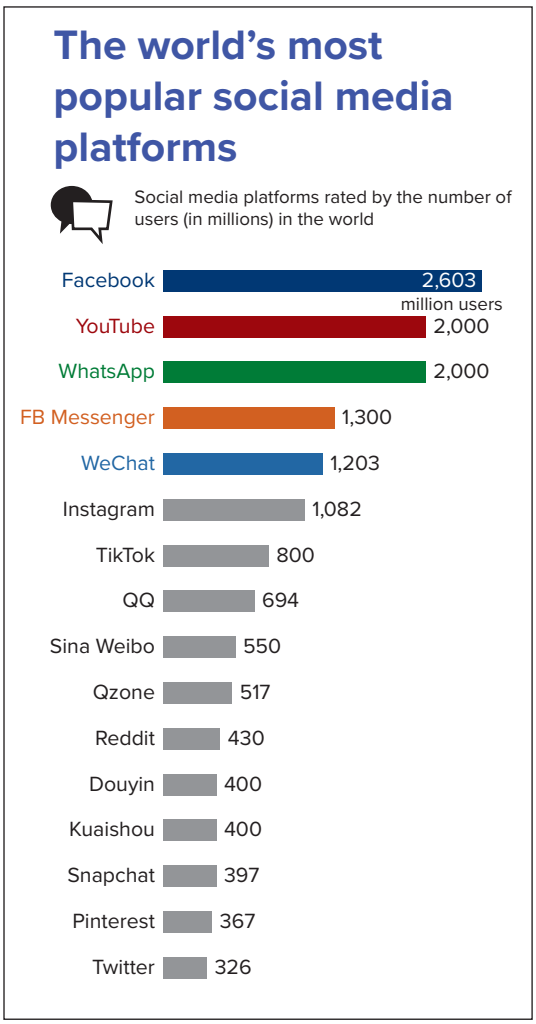
In addition to internal marketing data, vast amounts of external marketing data are available for companies to use in marketing analytics. Let's start with social media.

Data from Social Media

Each day, 3.78 billion people interact with social media such as **Facebook**, **Instagram**, **WhatsApp**, **TikTok**, **WeChat**, and **iMessage**,[6] In addition, each day 192 million **Twitter** users send 500 million tweets.[7] Exhibit 7.3 identifies the top social media sites based on the number of users.

Exhibit 7.3
Most Widely Used Social Media

(Source: Artem Minaev, "Internet Statistics 2021: Facts You Need-to-Know," First Site Guide, September 10, 2021, https://firstsite-guide.com/internet-stats/. Accessed 9/25/2021).



Users engage with social media to connect with friends, meet new people, stay current with news, express opinions, find a job, network, share videos and other content, and help identify products to buy, among other activities. Think of all the information available online through social media. For example, **Facebook** requires a name, birthday, and gender just to

[6]Source: Oberlo, "10 Social Media Statistics You Need to Know in 2021," 2021, https://www.oberlo.com/blog/social-media-marketing-statistics. Accessed 9/24/2021.
[7]Source: Oberlo, "10 Twitter Statistics Every Marker Should Know in 2021," 2021, https://www.oberlo.com/blog/twitter-statistics. Accessed 9/24/2021.

sign up—and that is before you add friends, photos, posts about your life, and likes/dislikes. In addition to using social media to advertise their products, companies can use **social media analytics (SMA)** to harvest and analyze user data from social media to make better-informed marketing decisions. SMA allows companies to see what customers are saying about their products and competitors' products; track keywords and hash tags; and analyze followers, activity, likes, and dislikes. SMA helps companies build user profiles and understand customer sentiment about a company's brand and products. In some cases, SMA is a part of a CRM system, providing a company with very rich data about its customers. Sometimes you will hear the term *digital (marketing) analytics,* which is a broad term for any analytics related to digital marketing. Thus, web analytics and social media analytics are examples of digital marketing.

Data About the Weather

Weather, which affects customer behavior and spending habits, is another type of external data that companies track and collect. For example, customers are willing to pay more for the same product when it is sunny. So, customers will pay $2.90 for a half gallon of orange juice when there is no sunlight, but $3.51 for the exact same orange juice when there is sunlight—a price increase of 21 percent! In addition, rain and a one-degree change in the temperature affect which products customers want to buy.[8]

Weather can also affect the success of marketing campaigns, for better or for worse, as in the following examples:

- **Michaels** increases advertising three days ahead of a rainy forecast because customers are more likely to purchase and engage in craft projects on rainy days.[9]
- To celebrate a new line of frozen treats, **Snapple** created a giant, 25-foot tall, 17.5-ton, strawberry-kiwi Snapple juice popsicle to break a world record during June in New York City. Not surprisingly, the popsicle melted, flooding Union Square and ruining the campaign.[10]
- **United Way** released a world-record number of balloons (1.5 million) in Cleveland, Ohio, right before a huge rainstorm. The rain pushed the balloons to the ground, which buried downtown in balloons, covered Lake Erie with balloons, and shut down an airport runway because of exploding helium balloons.[11] Even though United Way raised money ($1 sponsored two balloons), the balloonfest was a definite bust in other ways.

Data Regarding Culture, Laws, and the Business Environment

Culture may also impact marketing. Companies may need to create targeted messages because different languages, norms, and values can impact the success of marketing campaigns, as in the following examples:[12]

- In Italy, the campaign for **Schweppes** tonic water failed because the name translated to "Schweppes Toilet Water."
- In Spain, **Coors**'s campaign using the phrase "Turn it loose" failed because the phrase translated to "Suffer from diarrhea."

Companies also need to consider the laws and regulations of different countries when developing and marketing products. For example, Diet Coke is known as Coca-Cola Light

[8]Source: Kyle B. Murray, Fabrizio Di Muro, Adam Finn, and Peter Popkowski Leszczyc. (2010). "The Effect of Weather on Consumer Spending." *Journal of Retailing and Consumer Services, 17*(6), 512–520.

[9]Source: Weather Unlocked, *The Complete Guide to Weather Based Marketing,* http://www.weatherunlocked.com/media/1096/the-complete-guide-to-weather-based-marketing.pdf. Accessed 8/19/2021.

[10]Source: Bryan Smith, "Disaster on a Stick," 2005, https://www.nbcnews.com/id/wbna8321110. Accessed 8/19/2021.

[11]Source: Matthew Jarret, "1986 Cleveland Balloon Disaster," 1986, https://www.forgottenhistory.me/disaster/1986-cleveland-balloon-disaster. Accessed 8/21/2021.

[12]

paigns-that-failed-to-translate/. Accessed 4/01/2022.

in many non-U.S. countries because the word "diet" cannot be legally used in conjunction with Coke according to those countries' laws and regulations.

In addition, companies need data about the current state of the economy, business conditions, population trends, and customer demographics. A large amount of free data is available at the following websites:

- **https://www.data.gov/open-gov**: Contains free data about cities, states, and countries.
- **https://opendata.cityofnewyork.us**: Contains a treasure trove of data about New York City. Many other large cities have similar data sets available online.
- **https://www.census.gov/data.html**: Contains U.S. Census Bureau data about location and demographic data. Companies use these data to identify target markets and where to concentrate marketing efforts.
- **https://www.pewresearch.org**: The Pew Research Center provides free opinion research on a variety of topics.
- **https://www.cia.gov/the-world-factbook/**: The World Factbook provides a wide variety of information about the world's countries, including information about history, people, economy, and communications.
- Economic data are available from a variety of websites, including the websites of the U.S. Bureau of Economic Analysis, U.S. Small Business Administration, United Nations, and World Bank.[13]

Data from For-Profit Organizations

Several for-profit organizations sell information that companies can purchase to incorporate into marketing analytics. These organizations include:[14]

- **LexisNexis:** Provides legal resources, public records, and easy to search full-text articles.
- **Nielsen:** Uses representative panels around the world to determine what people are watching on the television and listening to on the radio. The data include time spent as well as demographic information about panel members. Companies can use these data to help decide where and when to spend advertising dollars to reach the target audience.
- **Statista:** Collects statistics for 170 industries in over 150 countries. The company provides historical statistics at the world, regional, and country level for narrowly defined industries, offering data on each industry's total sales revenue, average price, average revenue per capita, and sales volume. Companies can use these data to benchmark performance against the total industry. For example, **LEGO** would compare its performance to the entire Toy Construction Sets and Model industry.
- **Equifax, Experian, and TransUnion:** These three credit bureaus collect data about how quickly people pay off credit cards, bills, and loans. In addition, they pull data from public records about bankruptcies, tax liens, and other state and county data. Using these data, the bureaus create a personalized credit score, indicating a person's creditworthiness and ability to pay back debt. Companies can buy pre-screened data from the credit bureaus to select and pre-approve customers who want to make purchases on credit. The information they sell includes the applicant's age, name, gender, address, (unlisted) phone numbers, credit score, debt history, property information, and estimated income.

  A Google search will quickly bring you to each company's home page.

[13]The data listed are available at the following websites: https://www.bea.gov/, https://data.sba.gov/, http://data.un.org/, and https://data.worldbank.org/. Accessed 9/25/2021.

[14]The data listed are available at the following websites: https://www.lexisnexis.com/en-us/gateway.page, https://www.nielsen.com/us/en/, https://www.statista.com/, and https://www.experian.com/. The information in this paragraph is based on: Lumen, "Marketing Data Sources," https://courses.lumenlearning.com/marketing-spring2016/chapter/reading-marketing-data-sources/. Accessed 9/25/2021.

MINI CASE: IHOP's Burger Dilemma

IHOP (International House of Pancakes) had a problem. Even though it serves lunch and dinner, customers typically considered it a restaurant for breakfast, leading IHOP to make most of its sales early in the day. To increase sales revenue, IHOP wanted customers to come to the restaurant throughout the day for more than just pancakes. To achieve this goal, it introduced a new product, the Ultimate Steakburger, with a new marketing campaign. The campaign said that IHOP was changing its name to IHOB, short for International House of Burgers.

Marketing gurus deemed this campaign a huge failure. Social media blew up discussing the change (some for, some against). In jest, **Burger King** (temporarily) changed its name to the Pancake King on **Twitter**. IHOP eventually admitted that IHOB was just a marketing campaign, and that both the Ultimate Steakburger menu item and traditional IHOP name are here to stay.

Decision-Making with Data

How could IHOP use marketing analytics to determine if the IHOB marketing campaign was a success? Which internal and external data sources should IHOP have used to evaluate the success of the campaign?

Combining Internal and External Data

It is common for companies to use a wide variety of internal data and external data for marketing analytics. To assist with combining and analyzing heterogeneous data, companies might choose to use a data warehouse, data lake, or OLAP cube, as discussed in Chapter 2. Another way to pull everything together is to use customer journey analytics to link all of the interactions recorded in the company's CRM. A **customer journey** is visual representation of the experience that a customer has with a company, showing all of a customer's interactions or touchpoints with the company. **Customer journey analytics**, which tracks and analyzes how customers used different channels to interact with a company across time, helps companies determine which touchpoints are best at increasing revenue. They can then focus resources on those points.

The customer journey map varies from company to company. For example, **LEGO** uses an experience wheel. Unfortunately, LEGO will not share its true experience wheel, which it considers proprietary information. However, LEGO has created a representative experience wheel, shown in Exhibit 7.4, for a flight from London's Heathrow airport to New York City. The wheel shows the interactive points as well as the corresponding customer's (probable) emotional response. The wheel highlights areas that need improvement.

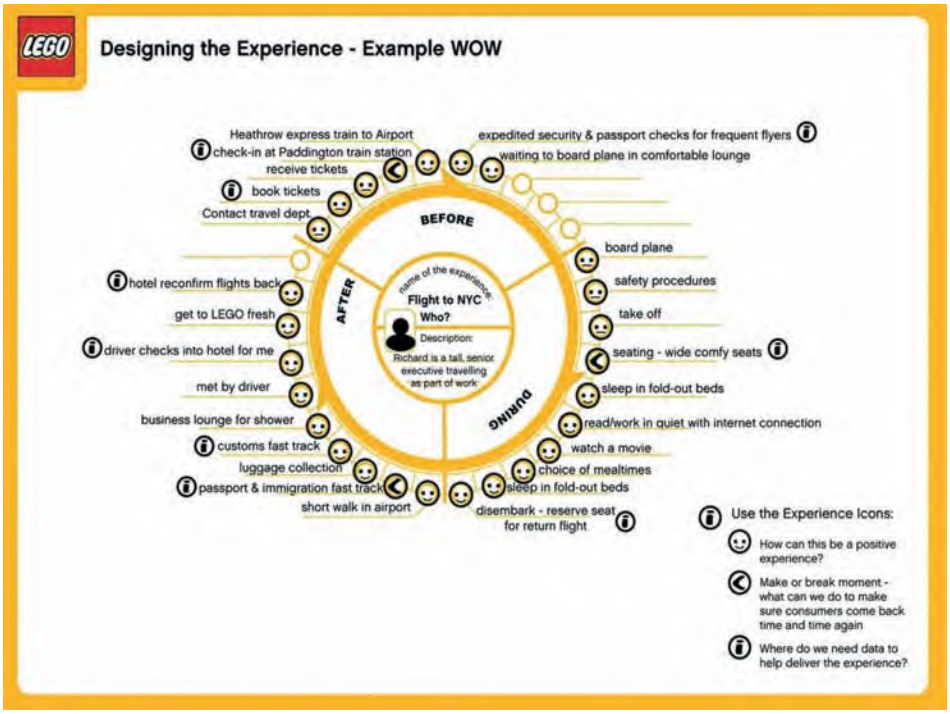


Regardless of how a company obtains or combines data, it must always be mindful of the customer's privacy. Data about customers are *overtly* captured when customers provide those data directly via emails, purchases, and online accounts. Data about customers are also *covertly* captured by a company's websites and phone apps, providing information about customers' page visits, location, contacts, and camera access. Some companies share these data with trading partners, marketing firms, data brokers, and government entities, which then use the data for purposes that customers did not agree to or that the company has no knowledge of.

For example, suppose a company's phone app collects GPS data, which are then sold to an auto insurance company to profile customers' driving habits—potentially affecting

Exhibit 7.4
Customer Journey Map: LEGO Experience Wheel Example

(Source: Act! Team, "4 Effective Customer Journey Map Examples," October 14, 2020, https://webolutions.com/experience-marketing-example-lego-creating-an-experience-with-intention/. Accessed 9/25/2021).



their auto insurance premiums. If customers discover the reason for the increased insurance premiums, they may be so angry that they refuse to purchase from the company again. To maintain customer loyalty and trust, and to conduct business ethically, it is important for companies to have good data governance and controls around all of their customer data.[15]



An Ethical Dilemma

The three credit bureaus (Equifax, Experian, and TransUnion) were set up to share information between lenders to evaluate and rate people's credit history. Creditors paid a fee to participate in the system, which helps them determine whether to extend credit. As part of this system, the credit bureaus collect a wide variety of information about people without their permission. More recently, the three credit bureaus have found a way to make extra money by selling the collected information. For example, when you apply for a new loan that requires a credit check, a credit bureau sells "triggered leads," tipping off competing lenders who can try to highjack the loan. Mailers for preapproved credit cards are also based on information sold from the credit bureaus. Some states, including California, are trying to limit this behavior. People can opt out from this process at www.optoutprescreen.com, but many people are not aware of this website. Do you think it is ethical for credit bureaus to sell your personal credit history to other companies to make money? Why or why not?[16]

[15]For more information, see Deloitte, "Five Questions About the Misuse of Consumer Data," 2019. Available online at https://www2.deloitte.com/us/en/pages/advisory/articles/five-questions-about-the-misuse-consumer-data.html. Accessed 1/15/2022.

[16]Source: Jeff Lazerson, "Should Credit Bureaus Sell Your Data When You Apply for a Mortgage?" *The Orange County Register,* November 14, 2019. Available online at https://www.ocregister.com/2019/11/14/should-credit-bureaus-sell-your-data-when-you-apply-for-a-mortgage/. Accessed 4/01/2022.

PROGRESS CHECK

6. Identify sources of internal data that can be used for marketing analytics.
7. Identify sources of external data that can be used for marketing analytics.
8. How can companies combine various sources of marketing data to gain insight and inform marketing decisions?
9. Based on LEGO's experience wheel in Exhibit 7.4, what areas need to be changed to improve the customer experience? What type of data should the company collect in the interactions flagged with the "need data" icon?
10. Which types of data sources might be the most important for LEGO? Explain.

7.4 DESCRIPTIVE MARKETING ANALYTICS

LO 7.4

Describe and provide examples of descriptive analytics used to answer marketing questions.

After specifying the marketing questions and obtaining appropriate data, companies are ready to perform marketing analytics. In the rest of this chapter, we focus on four types of marketing analytics: descriptive, diagnostic, predictive, and prescriptive. While a variety of statistical software tools, including R and SAS, can be used to analyze the data, we continue to focus on Excel, Tableau, and Power BI.

Descriptive analytics describes or summarizes the past, highlighting the characteristics and features of a data set by summarizing a large amount of data in an easy-to-understand format. It does not include any assumptions or future estimates/predictions. In other words, descriptive analytics helps us understand the *actual* data and what has already happened.

Descriptive analytics uses a wide variety of summarization and statistical techniques as well as simple visualizations. We discuss some examples in the following sections.

Statistical, Summarization, and Data Visualization Techniques for Descriptive Analytics

Data analysts use summarization and statistical techniques to answer the questions "What happened?" or "What is happening?" These techniques characterize, summarize, and organize data and are generally available in spreadsheet software such as Excel. For example, counts, totals, sums, averages, subtotals, minimums, maximums, pivot tables, standard deviations, and ratios are all descriptive analytics that Excel can generate. In addition, business analysts can use Excel to create simple graphs/visualizations such as histograms, bar charts, and pie charts. Exhibit 7.5 provides examples of statistical techniques that can be used to answer questions about the different components of the marketing mix.

Examples of Marketing Descriptive Analytics

Descriptive analytics provides the basic statistics for marketing data and the marketing mix. For example, the **Rebrickable** website (rebrickable.com) tells you how to "combine your old LEGO to build new creations." It also provides a downloadable database of information on LEGO sets since 1949. We can use these data to examine LEGO's *product* offerings. Specifically, let's use this data set to answer the question "How many LEGO sets does LEGO release a year?"

Our question indicates that the desired level of analysis is LEGO sets by year. Exhibit 7.6(a) shows a sample of the data from Rebrickable. The file consists of five columns/variables: unique set id (*set_num*), unique set name (*name*), year of release (*year*), theme (*theme_id*), and number of parts in the set (*part_num*). Each row represents one LEGO set. There are 18,393 rows in the file, indicating that LEGO has created/released 18,393 LEGO sets (not shown).

Statistical Techniques for Descriptive Analytics	Purpose	Example Marketing Questions
Counts	Determine frequency of occurrence	How many people visited our website? How many website visitors did not purchase any products? How many website visitors purchased at least one product? How many new Twitter followers did we get as a result of our last post?
Graphs: bar charts, histograms	Provide summary visualization	How much sales revenue for a product was generated in each marketing channel in 202X? How many units were sold by each salesperson in 202X?
Minimums, maximums, medians, standard deviations	Provide summary measures of dispersion	What are the top 10 products in terms of sales and profit in 202X? How much does sales revenue vary by salesperson?
Pivot tables	Provide a flexible way to summarize large amounts of data	What were total sales for each month in 202X? What were total sales by salesperson?
Ratios	Compare two numbers	What is the company website bounce rate in January (that is, how many website visitors enter the company website and immediately leave)? What is the conversion rate for the company website in May?
Totals, sums, averages, subtotals	Provide summary measures of performance	What was the average monthly sales revenue from a product in 202X? What was the average price of the product in 202X? How many total product units were sold in June? In each region? In each store? In each marketing channel?

Exhibit 7.5
Statistical Techniques for Marketing Descriptive Analytics

Before we can analyze the data, we need to transform it into the proper format. Let's first determine the total number of sets released each year. To do so we must summarize the Rebrickable data by year. Using a pivot table, we can count the number of sets released each year from 1949 to 2021. Let's call this variable, *Num_Yearly_Sets.* Exhibit 7.6(b) shows a sample of the pivot table for years 1949 through 1958. Each year has its own row for *Num_Yearly_Sets.* For example, in 1949 LEGO released five LEGO sets. Given the 71 years of data, the pivot table has 71 rows for *Num_Yearly_Sets,* one for each year. While not shown, if you add up the 71 *Num_Yearly_Sets* values, the sum equals the total number of LEGO sets released: 18,393.

Now, using the Descriptive Statistics menu option of the Analysis ToolPak for Excel (see Appendix E for more information), we can easily answer the question by generating descriptive statistics on *Num_Yearly_Sets* between 1949 and 2021. Exhibit 7.6(c) shows

(a) Sample of Original Data

set_num	name	year	theme_id	num_parts
700.1-1	Extra-Large Gift Set (ABB)	1949	365	142
700.2-1	Large Gift Set (ABB)	1949	365	178
700.3-1	Medium Gift Set (ABB)	1949	365	142
700.A-1	Small Brick Set (ABB)	1949	371	24
700.B-1	Small Doors and Windows Set (ABB)	1949	371	12
700.1.1-1	Single 2 x 4 Brick (ABB)	1950	371	1
700.1.2-1	Single 2 x 2 Brick (ABB)	1950	371	1

(b) Sample of Pivot Table

Row Labels	*Num_Yearly_Sets*
1949	5
1950	6
1953	4
1954	14
1955	29
1956	13
1957	20
1958	49

(c) Num_Yearly_Sets

Mean	259.06
Standard Error	34.09
Median	114.00
Mode	39.00
Standard Deviation	287.25
Sample Variance	82,510.88
Kurtosis	0.28
Skewness	1.19
Range	1,063.00
Minimum	3.00
Maximum	1,066.00
Sum	18,393.00
Count	71.00

Exhibit 7.6
Summary Statistics for LEGO Set Releases, 1949–2021.

(Source: https://rebrickable.com/downloads/, sets.csv.gz, Accessed 1/07/2022.)

the mean, median, mode, standard deviation, and other summary statistics. According to the descriptive analytics, LEGO has released 18,393 LEGO sets (Sum of *Num_Yearly_Sets*) in the 71-year period analyzed (Count of *Num_Yearly_Sets*). In addition, the number of annual LEGO sets launched has varied greatly, ranging from 3 sets to 1,066 sets, with an average (mean) of 259 new LEGO sets released each year.

If we want to create a data visualization of our results, we can use Tableau. For example, we can easily generate a horizontal bar chart to answer the question "Which LEGO theme had the most LEGO sets released from 1949 to 2021?" Exhibit 7.7 depicts a Tableau visualization that combines two Rebrickable tables (sets and themes) to generate the bar chart. Based on the chart, the themes with the most set releases are Gear, Star Wars, and Technic.



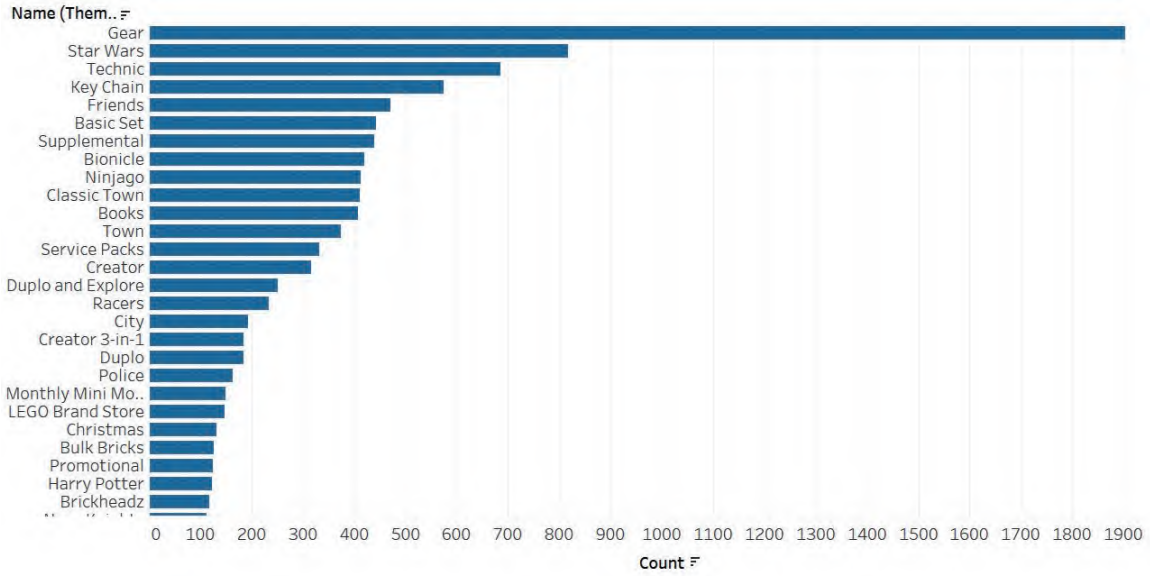


Exhibit 7.7
Horizontal Bar Chart for LEGO Set Releases by Theme, 2021

(Source: https://rebrickable.com/downloads/, sets.csv.gz and themes.csv.gz. Accessed 10/23/2021.)

LAB CONNECTIONS

LABS 7.1, **7.2**, and **7.3** provide step-by-step instructions on how to use different features in Excel, Power BI, and Tableau to generate descriptive analytics.

- **LAB 7.1** shows how to create descriptive statistics, a bar chart, and a histogram in Excel.
- **LAB 7.2** shows how to create descriptive statistics in Excel using pivot tables.
- **LAB 7.3** shows how to create a histogram in Tableau and Power BI.

Regarding the *place* of the marketing mix, we can answer the question "How did the number of LEGO stores change in the first half of 2021?" Exhibit 7.8 shows what LEGO Group reported in the summary of its first half of 2021 financial results. In the first half of 2021, LEGO opened 60 new stores around the world, bringing the total to 737 LEGO stores. LEGO could have also stated that the number of stores grew by 9 percent (60/(737-60) * 100) in the first half of 2021. Exhibit 7.8 shows that LEGO is focusing on developing a digitally enabled brand as well as a memorable brand experience with a new flagship store in New York City.

Exhibit 7.8
Summary Statistics for LEGO, 2021

(Source: The LEGO Group 2021 Financial Results, https://www.LEGO.com/en-us/aboutus/news/2021/september/2021-interim-results. Accessed 10/23/2021.)



Ratios: Key Performance Indicators (KPIs)

"If you don't *measure* it, you can't *improve* it."
—CHRIS GETMAN, Vital

Key performance indicators (KPIs) are numerical metrics used to gauge company performance. They are an integral part of descriptive and diagnostic analytics for marketing. The possible number of KPIs is endless. Each company determines which KPIs are the most crucial for the business and how they should be broken down—for example, by channel, product, region, or store. The following KPIs are commonly used in marketing.[17]

- **Sales revenue.** This KPI can measure sales in total, on average, per visit, and/or as a percent change from a prior period.
- **Leads.** Leads are potential customers. They can be divided into various groups:
  - *Marketing qualified leads (MQLs)* are consumers who expressed some interest in the product—for example, by downloading a white paper from the company's website.
  - *Sales qualified leads (SQLs)* are consumers who are more likely to buy a product/service and are therefore worth pursuing.
  - *Lead-to-close ratio* or *lead-to-customer ratio* is the number of leads buying a product/service divided by number of leads received.
- **Customer lifetime value.** This KPI is the value of a customer over the customer's lifetime. It can be based on sales revenue or gross profit margin (that is, margin after costs[18]). We prefer the gross margin approach:

$$\text{Customer lifetime value} = \text{Sales revenue} * \text{Gross profit margin} * \text{Average number of repeat purchases}$$

- **Customer acquisition cost.** This cost can be calculated per visit, per sale, and/or per customer.

$$\text{Customer acquisition cost} = \frac{\text{Total marketing investment}}{\text{Customers acquired}}$$

[17]This section is based on Ryan Stewart, *31 Marketing KPIs to Use in Your Campaign,* July 29, 2019, https://webris.org/marketing-kpis/. Accessed 10/23/2021.
[18]Some companies include customer acquisition costs.

- **Return on marketing investment (ROMI).** This KPI indicates the rate at which marketing spending contributes to profits. It measures the company's effectiveness in using the resources allocated to marketing efforts.

$$\text{ROMI} = \frac{(\text{Sales} * \text{Gross margin \%}) - \text{Marketing expenditures}}{\text{Marketing expenditures}}$$

MINI CASE: Working with Barilliance E-Commerce Analytics

Exhibit 7.9 presents sample KPIs generated by an e-commerce analytics tool called **Barilliance** over the last seven days for a company. The exhibit shows that the company's website had 227,483 total visitors; 20,335 people who placed items in their shopping cart abandoned it, choosing not to purchase over $1.15 million in products. The exhibit also shows that 47 percent of emails sent were opened, leading to 349 sales. Based on the data in the exhibit, we calculate the overall conversion rate as less than 2 percent (= 349/20,307 * 100).

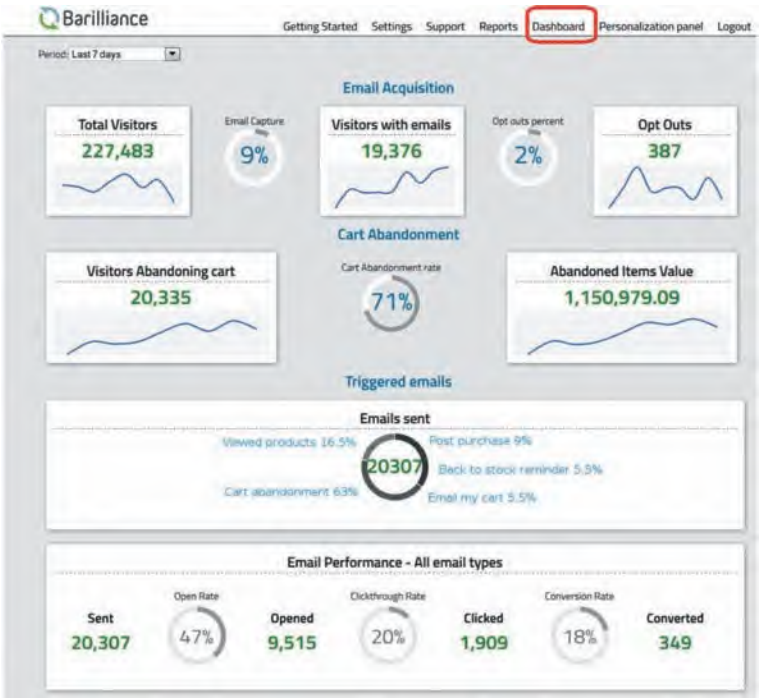


Exhibit 7.9
Sample KPIs by Barilliance

(Source: https://www.barilliance.com/cart-abandonment-rate-statistics/. Accessed 9/25/2021.)

Decision-Making with Data

What does the Barilliance analysis tell you about this company's effectiveness in terms of sending emails to customers? Based on the analysis, what would you recommend to the company? What additional analysis would you like to see?

- **Website information.** Companies can gather information about website traffic (number of visitors), page views, session duration, **bounce rate** (percentage of website visitors who enter the company website and immediately leave), cart abandonment rate (percentage of website users who put product in the cart but do not buy), conversion rate (percentage of potential customers who actually purchase products), followers, new visitors, and keywords used by customers to find a product/brand/company from web search engines such as **Google** and **DuckDuckGo**.
- **Year-over-year (YOY) growth.** YOY evaluates a company's progress over time. It compares a KPI from one period against the KPI from the same period of the prior year. It helps mitigate seasonality that may affect KPIs.

$$\text{YOY} = \frac{(\text{This year's KPI} - \text{Last Year's KPI})}{\text{Last Year's KPI}} * 100$$

- **Churn.** Churn is the percentage of customers that stopped using a company's service or product, typically measured during a certain time frame.
- **Other information.** Companies can gather information about newsletter/email open and click-through rates, the number of calls from potential customers, likes/dislikes, product/service quotes, and store traffic.

PROGRESS CHECK

11. How are bounce rate and conversion rate related? How can marketers reduce the bounce rate?
12. For LEGO, which KPIs do you think are the most important? Explain.

LO 7.5

Describe and provide examples of diagnostic analytics used to answer marketing questions.

7.5 DIAGNOSTIC MARKETING ANALYTICS

"Eliminate all other factors, and the one which remains must be the truth."

—**SHERLOCK HOLMES**, in Arthur Conan Doyle, *The Sign of the Four* (1890)

Diagnostic analytics helps to answer the questions: "Why did it happen?" and "What are the causes of past results?" In other words, diagnostic analytics attempts to identify the root cause or the factor that initiated the chain of events. Descriptive analytics usually cannot uncover this information. Diagnostic analytics helps decision-makers evaluate the relations, patterns, and linkages between variables that might be related to an anomaly. Performing diagnostic analytics often requires combining multiple data sets and sometimes using both external and internal data.

Identification of Anomalies and Outliers

Often the first step in diagnostic analytics is to look for and identify the unusual, unexpected results known as anomalies or outliers. Exhibit 7.10 details some anomalies for the marketing mix components and provides some possible explanations.

After pinpointing the unexpected results, a company can investigate to understand exactly why those results occurred. Root cause analysis is often used to understand what is driving an outlier or anomaly.

Root Cause Analysis

Root cause analysis is the process of determining what factor, or factors, caused a change. The root cause can drive the change itself or trigger a series of other events that eventually result in the change. There may be more than one root cause; root cause analysis focuses on the internal

Marketing Mix Component	Expectation	Example of Anomaly	Possible Explanations for the Anomaly
Product	Our most popular product will continue to sell well.	Product sales are dropping.	A competitor introduced a similar product with more features for about the same price. A recession reduced customers' discretionary income.
Price	Market will grow over time for a new product.	Market growth rate is low (or almost nonexistent).	Price is too high. A lower price is needed to attract new customers. A competitor introduced a new product at a lower price.
Place	Sales of a product will be the same at all stores.	One store suddenly sells less product than other stores.	A new mall was built nearby. As a result, traffic at this store has decreased. Road construction made getting to the store difficult.
Promotion	Website revenue will continue to grow.	Website revenue is down 10 percent.	Promotional email click rate and conversion rate have dropped. Website is not functioning properly due to denial-of-service attack or lack of functionality.

Exhibit 7.10
Diagnostic Analytics: Examples of Marketing Anomalies for Components of the Marketing Mix

and external factors with the highest probability of causing the change. If the anomaly has a negative impact on the company, then the root cause needs to be fixed. If the anomaly has a positive impact on the company, then the root cause needs to be leveraged across the company.

Root cause analysis focuses on factors that can be placed into four groups:

- The root cause itself
- Unrelated factors
- Correlated factors, which are other symptoms of the root cause
- Contributing factors in the chain of events


DATA VISUALIZATION

One method of identifying the root cause places suspected factors in a timeline, with the understanding that the cause of the anomaly/outlier could have happened days, weeks, or months before the change, especially if it involved a chain of events. A second method is a visualization technique called an **Ishikawa diagram**, also known as a *fishbone, herringbone,* or *cause-and-effect diagram.* The Ishikawa diagram captures the underlying factors (and the relationships) that may be the root cause of the problem. It should include all possible causes, including unrelated, correlated, contributing, and root. Exhibit 7.11 is an Ishikawa diagram listing the primary factors that could explain why website revenue is down 40 percent. It is categorized by the four traditional Ps as well as process, which captures KPIs about the customer's interaction with the website.

Using Exhibit 7.11's Ishikawa diagram as a guide, we can analyze the listed factors to uncover additional details hidden in the data. Techniques to reveal the hidden information include summarizing the data at different levels as well as **drilling down**, **drilling up**, and **drilling through**, which provide more, fewer, and alternative views of the data, respectively. The goal is to find patterns or linkages in the underlying data set or with an external data set.

SOARing to Success

Understanding the Use of KPIs in Data Analytics

The KPIs introduced in the descriptive analytics section of this chapter are often also used in diagnostic analytics. The difference is that descriptive analytics calculates a KPI at one point in time, while diagnostic analytics examines how the KPI changed over time.

Exhibit 7.11
Ishikawa Diagram Example

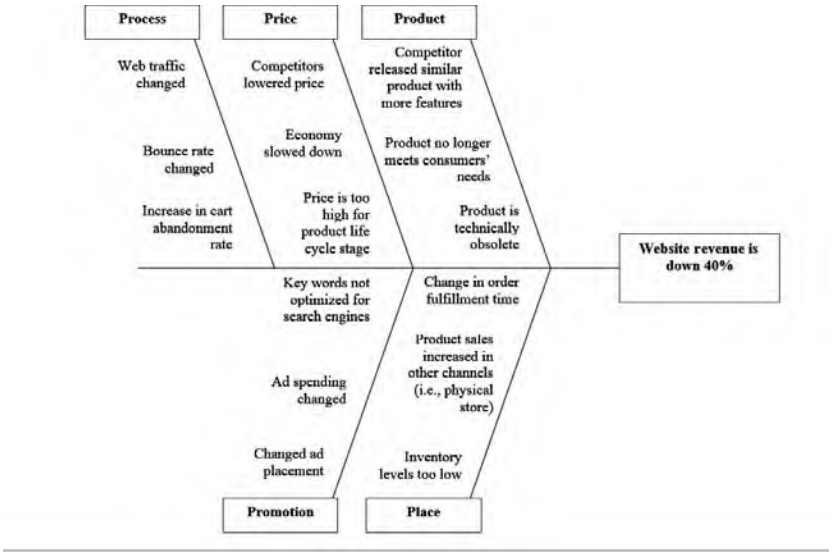


Exhibit 7.12 summarizes the diagnostic analytics techniques that can be used in marketing and examples of marketing questions that can be answered.

Statistical Technique for Diagnostic Analytics	Use	Example Marketing Question
Cluster analysis	Groups observations into clusters/groups where the data within any cluster are similar.	Which customers are similar in terms of purchase frequency, purchase amounts, or purchase channel?
Correlation	Determines the extent to which variables are related to each other.	To what extent are promotional emails related to sales revenue?
Hypothesis testing	Tests whether variables are significantly different from each other.	Has website traffic significantly dropped between two time periods?
Pivot table/ cross-tabulation	Groups variables to understand the relationships between them. Allows drill down and drill up.	What is the future purchasing behavior of customers who buy in the same month (cohort analysis)?
Regression	Examines whether there is a relationship between an input and an outcome.	How is advertising expense related to sales revenue?
Sales funnel chart	Shows the conversion rate at each step in the sales process using an inverted pyramid visualization displaying the number of (potential) customers at each step.	What percent of customers who ask for a quote on a product eventually make a purchase?
Summarization	Shows how totals vary at different levels of granularity.	How many website visitors actually make a purchase?

Exhibit 7.12
Diagnostic Analytics Techniques for Marketing

Cluster Analysis[19]

Cluster analysis is a statistical technique that organizes data into clusters, or groups, in such a manner that all of the data within any cluster are similar, while data across different clusters are different. Technically speaking, cluster analysis minimizes the difference

[19]This analysis is roughly based on the analysis found at https://select-statistics.co.uk/blog/customer-segmentation/. Accessed 10/23/2021.

between each data point and the center of each cluster, and data points closest to each other form a cluster. Cluster analysis relies on an unsupervised, learning algorithm in which the company does not have predefined categories.

In marketing, cluster analysis can be used for market segmentation, which divides the target market into different groups. Market segmentation allows a company to better understand each group's needs and develop a targeted (group-specific) marketing initiative with maximum impact.

For example, let's assume that **LEGO** surveyed its customers and collected data to capture the following two dimensions, each on a scale of 0 to 5:

(1) *price conscious score,* which is how price affects the customer's intention to buy a LEGO set (0 means the customer does not care about price and 5 means the customer is interested only if the price is low), and
(2) *LEGO loyalty score,* which is how the customer feels about the LEGO brand (0 means the customer has no loyalty and 5 means the customer loves the brand.)

Using these hypothetical data, Tableau's cluster analysis creates three groups as shown in Exhibit 7.13. Group 1, the blue cluster, is very price conscious and not loyal to the LEGO brand. Customers in this cluster will only buy an inexpensive LEGO set and would probably buy a competitor's inexpensive product. Group 2, the orange cluster, is medium on both price consciousness and LEGO loyalty. These customers generally like the LEGO brand, but they are still concerned about price. Group 3, the red cluster, is composed of dedicated LEGO lovers, who are low on price consciousness and high on LEGO loyalty. LEGO can use the results to develop different marketing initiatives for the three different groups. Given this scenario, we're guessing that LEGO would try to convert Group 2 customers into LEGO lovers!

Note: Exhibit 7.13 is an easy example to interpret, but most cluster analyses are not as straightforward.



Exhibit 7.13 Diagnostic Analysis in Tableau Using Customer Clusters

Tableau Software

Correlation and Summarization[20]

Let's assume that LEGO has experienced a drop in sales from its website. Using this scenario, we can use diagnostic analytics techniques to answer the question "Why has sales revenue dropped for the company's website?"

[20]The content in this section was inspired by Jeff Rajeck, "Analytics Approaches Every Marketer Should Know #2," *Diagnostic Analytics,* July 31, 2017, https://econsultancy.com/analytics-approaches-every-marketer-should-know-2-diagnostic-analytics/. Accessed 9/25/2021. Unfortunately, real daily sales revenue and web traffic numbers for LEGO are not available to us. We used the monthly traffic figure and bounce rate for LEGO.com, based on **Similarweb**'s statistics, to guide our daily web traffic numbers and the average cost of a LEGO set to simulate web sales.

After consulting the Ishikawa diagram in Exhibit 7.11, we decide to investigate web traffic. Using the CORREL function in Excel, we find that the relationship between (hypothetical) sales revenue and (hypothetical) web traffic is highly and positively correlated at a value of 0.88, as shown in Exhibit 7.14. The higher the traffic, the higher the sales revenue. Correlation does not always mean causation, however.

Exhibit 7.14
Correlation Analysis of Web Traffic and Sales Revenue in Excel

Sample Data		
Day	Traffic	Sales Revenue
1/1/2022	765,000	$27,348,750
1/2/2022	742,050	$28,197,900
1/3/2022	803,250	$32,130,000

Correlation Analysis

=CORREL(B2:B425, C2:C425)

=0.88

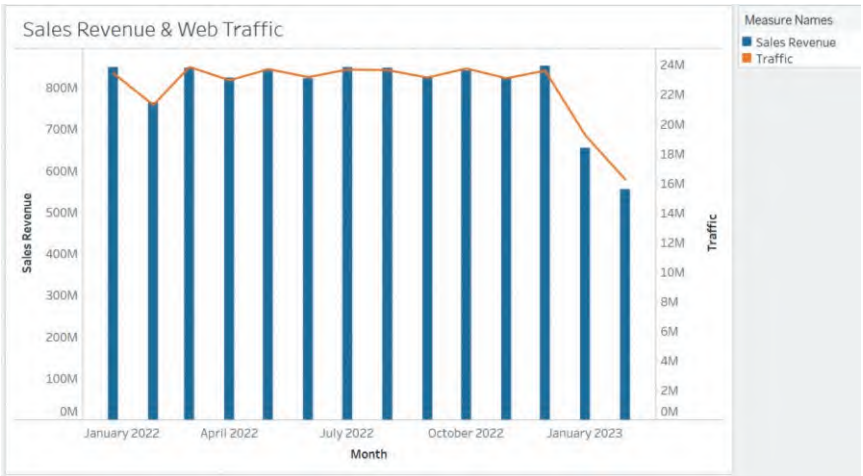


DATA VISUALIZATION

We can use Tableau to graph sales revenue and web traffic, both summarized by month. As Exhibit 7.15 shows, both figures dropped after December 2022.

Exhibit 7.15
Diagnostic Analysis in Tableau Summarizing Sales Revenue and Web Traffic by Month

Tableau Software



Continuing to drill down, we compare daily figures from December 2022 to February 2023 using Exhibit 7.16, which shows that on January 7 traffic to the LEGO website dropped sharply. Further analysis reveals that on January 7, the process used by search engines such as **Google** to determine which websites should be returned from a user's search changed, causing the drop in web traffic to the LEGO website. To solve the problem, LEGO needs to update its keywords so that they are optimized for search engines.

Sales Funnel Chart



DATA VISUALIZATION

A **sales funnel chart** is an inverted pyramid chart used to show the steps in the sales process. The sales funnel chart is the widest at the top because it represents the largest pool of potential customers. Each successive step is smaller, with the bottom step the smallest because it represents the people who actually purchased a product or service from the company (the customers). In other words, a sales funnel chart is one way to depict a customer's journey through the sales steps. The size of each step in the funnel reflects the relative numbers of (potential) customers and potential sales revenue at each step.

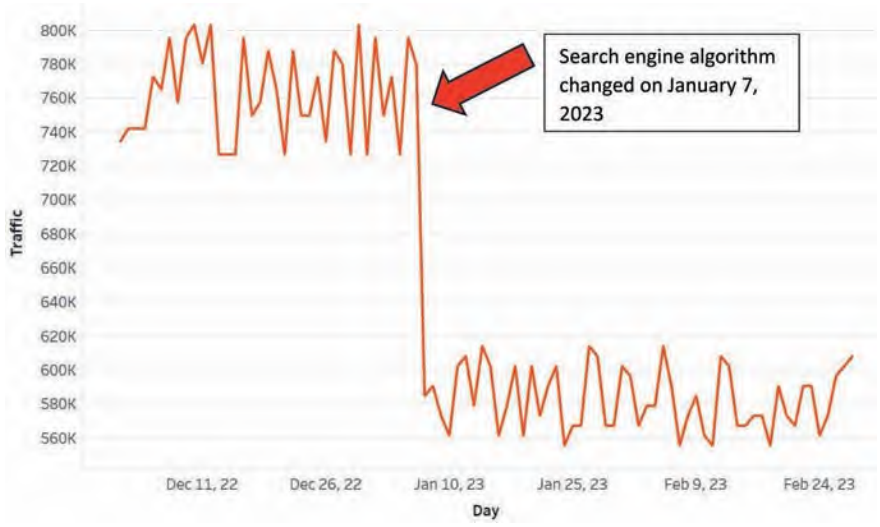


Exhibit 7.16
Diagnostic Analysis in Tableau Using Daily Website Traffic Data
Tableau Software

Looking at the LEGO website (www.LEGO.com/en-us), it appears that LEGO has the following steps in its sales process:

1. Potential customers land on LEGO.com.
2. Potential customers select "Shop."
3. Potential customers select a subcategory of Shop (for example, Age, Clothing, Sets by Theme).
4. Potential customers place one or more items in the shopping cart.
5. Customers pay for items in the shopping cart and complete the purchase.

Based on these five steps in the sales process, LEGO's sales funnel chart may look something like Exhibit 7.17. LEGO could customize the sales funnel chart by adding percentages and/or actual numbers for each step.

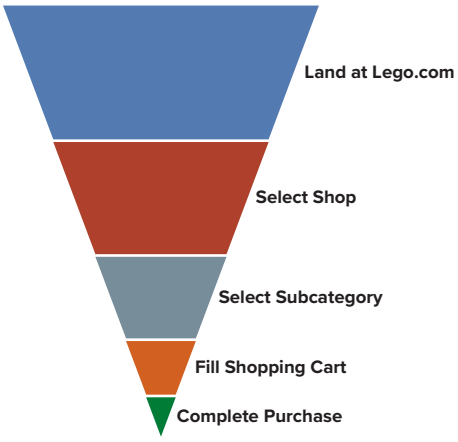


Exhibit 7.17
Diagnostic Analysis Using a Sales Funnel Chart

LAB CONNECTIONS

LABS 7.4 and **7.5** provide step-by-step instructions on how to use diagnostic techniques to answer marketing questions.

- **LAB 7.4** uses Excel to analyze website data using a sales funnel chart.
- **LAB 7.5** uses Tableau to show how to group customers into clusters.

PROGRESS CHECK

13. What is the difference between drill down, drill up, and drill through? If a company wants to look at data in a different manner than currently presented but at the same summarization level, which method would be appropriate: drill down, drill up, or drill through?
14. Explain how you would use additional diagnostic techniques to further analyze LEGO's (fictional) drop in website traffic as summarized in Exhibit 7.16.

LO 7.6

Describe and provide examples of predictive analytics used to answer marketing questions.

7.6 PREDICTIVE MARKETING ANALYTICS

Predictive analytics answers the questions: "Will it happen in the future?" and "What is the probability something will happen?" and "Can we forecast what will happen?" The goal is to predict future performance using historical data and a user-specified **algorithm**, which is a set of instructions on how to combine data. Unlike descriptive and diagnostic analytics, which describes things we know are true, predictive analytics describes things that possibly might happen.

There are three broad categories of predictive analytic techniques: (1) classification, (2) decision tree, and (3) regression.[21] Exhibit 7.18 outlines these three predictive analytics techniques, summarizes how they are used, and provides an example of a marketing question that each technique can help answer.

Statistical Technique for Predictive Analytics	Use	Example Marketing Question
Classification	Separates a data set into two or more predefined "like" groups.	Which customers are more likely to buy again?
Decision tree	Predicts a specific dependent (outcome) variable based on a pathway that includes all action points with associated probabilities.	Which marketing activities will generate more sales?
Regression	Predicts an outcome based on the relationship between independent and dependent variables.	What will sales be if advertising expense is increased?

Exhibit 7.18
Predictive Analytics Techniques for Marketing

[21] Decision trees can be used to create classification trees (for categorical data) and regression trees (for continuous data).

Decision Tree


DATA VISUALIZATION

A **decision tree** is a decision-support tool that visually shows various outcomes from different (sequential) decisions. Each branch represents different decisions. The decision tree may also include probabilities and costs of different paths, allowing expected value calculations. Decision trees are used to help develop a strategy to achieve a desired outcome. To be effective, all possible pathways need to be included and each pathway should be *mutually exclusive.* That is, if one path is selected, another path cannot also be used.

Decision trees have many benefits: (1) They are easy to understand, visualize, and interpret. (2) They can handle nonlinear relationships that may be difficult to analyze using other techniques. (3) They require minimal data prep. In addition, decision trees can be used by all different business functions in many different contexts. Here we focus on a marketing example of a decision tree.

Assume that **LEGO** is trying to answer the question "Should we expand our product offerings, or focus on strengthening sales of existing products?" Exhibit 7.19 presents the hypothetical decision tree for this question. The top branch shows the decision to expand the product line. The bottom branch shows the decision to focus on strengthening existing product sales. The tree diagram also shows the potential outcomes, associated probabilities, and net profit. For example, designing new products can result in zero new products, one new product, or many new products with projected net profits (in millions) of – $40, $105, and $200 with 45 percent, 40 percent, and 15 percent probability, respectively.

Using this information, the expected value of this decision is approximately $54 million ((0.45 * – $40) + (0.40 * $105) + (0.15 * $200) = $54.40). Similarly, the expected value of strengthening product lines is approximately $43 million ((0.75 * $60) + (0.25 * – $10) = $42.50). Based purely on expected value, the company should choose to design new products because $54 million is greater than $43 million. However, other factors, such as a shortage of LEGO designers or management's long-term vision, may be considered when deciding which option to pursue, leading to a different decision.

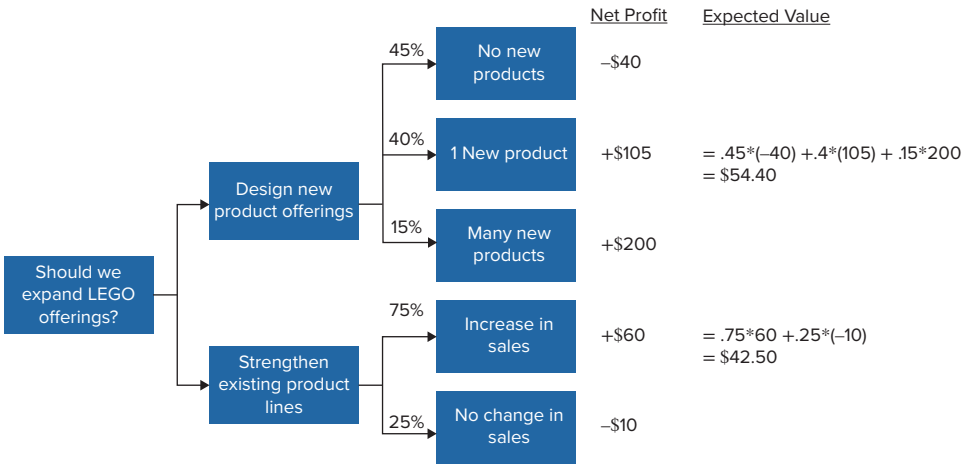


Exhibit 7.19
Predictive Analytics Decision Tree Example

Regression

Chapter 3 introduced and explained regression in detail. In predictive analytics, we can use the regression model to help predict a given output variable. That output variable is the dependent/outcome variable (the y), as it is the variable that you are trying to predict based

on a variety of input variables. Those input variables are the independent variables (the x), as they are the variables that you hypothesize relate to the dependent (outcome) variable and can help explain or predict its outcome. The regression equation is: $y = b + mx$, where b is the intercept, y is the dependent (outcome) variable, x is the independent variable, and m is equal to the slope of the line. If you need a refresher on regression, please review Chapter 3.

Companies use regression to predict the relationship between variables that relate to marketing. Here are a few examples, with the dependent/outcome variable (y) first and the independent variable (x) second:

- Sales revenue and advertising expense
- Email signups and advertising expense
- Conversion rates and advertising expense
- Sales revenue and internet advertising pay-per-click cost
- Customer satisfaction and product features
- Customer loyalty and product features
- Customer satisfaction and product availability
- Sales revenue for another complementary product and expenditures for a related product
- Sales revenue and free shipping

Regression can also help answer the question "Which social media outlet has the largest impact on sales revenue?" This regression would include all of the social media outlets used by the company as independent variables and sales revenue as the dependent variable. The results of this regression can be a powerful tool helping companies understand sales drivers to focus their marketing resources in the proper places.

LAB CONNECTIONS

LAB 7.6 provides step-by-step instructions on how to use regression to predict sales revenue from advertising expense.

- **LAB 7.6 Excel** provides instructions on how to do regression in Excel.
- **LAB 7.6 Tableau** provides instructions on how to add a trend line to a scatterplot in Tableau.

PROGRESS CHECK

15. Which predictive analytics technique should be used to determine how a change in customer satisfaction will affect sales revenue?
16. Generate additional marketing-focused questions that can be answered by the three predictive analytics techniques summarized in Exhibit 7.18.

LO 7.7

Describe and provide examples of prescriptive analytics used to answer marketing questions.

7.7 PRESCRIPTIVE MARKETING ANALYTICS

Prescriptive analytics answers the questions: "What should we do, based on what we expect will happen?" and "How do we optimize our performance based on potential constraints?" While both predictive and prescriptive analytics make predictions about the future, prescriptive analytics also recommends a course of action. Exhibit 7.20 summarizes some prescriptive analytics techniques and provides examples of marketing-focused questions that

each technique could answer.[22] We already saw some of these techniques, including goal-seek analysis, scenario analysis, and sensitivity analysis, in earlier chapters.

Statistical Technique for Prescriptive Analytics	Use	Example Marketing Question
Cash flow analysis	Evaluates future cash flows for potential investments or expenditures. It typically uses net present value (NPV) or internal rate of return (IRR)	In which marketing campaign should the company invest?
Goal-seek analysis	A type of what-if analysis that determines which input value(s) are needed to achieve a desired output (that is, backward analysis)	How much improvement in conversion rate is needed to sell 100,000 units?
Optimization	The process of making something as effective as possible	Which social media platform's users respond best to ads?
Scenario analysis	A type of what-if analysis that examines the effect of changing multiple input variables at the same time on an outcome variable (typically, base/most likely case, worst case, and best case)	What is the number of products sold given the worst case for projected changes in the overall economy?
Sensitivity analysis	A type of what-if analysis that examines the effect of changing one input on an outcome variable	What happens to forecasted net income if units sold increase by 25 percent?
Sentiment analysis	A technique that determines the emotion behind words	What marketing campaign should the company use to generate positive customer reviews?

Exhibit 7.20
Prescriptive Analytics Techniques and Examples

Goal-Seek Analysis

Digital marketing includes placing advertisements on search engines or carefully selected websites that are frequently visited by a company's target customers. To place an advertisement on a search engine or website, companies pay a fee. These fees are sources of revenue for search engines and websites.

Fees can be charged based on the number of clicks or the number of views. **Cost-per-click (CPC) fees** are based on the number of clicks. In this fee structure, a company is charged only when a user clicks on the ad. For example, if the CPC is $0.25 per click and 1,000 users click on the ad, then the company will owe $250 to the sponsoring website. **Cost-per-mille (CPM) fees** are based on the number of impressions or views, counted in the thousands (*mille* means 1,000 in Latin). For example, if the CPM is $0.15 per view and 1,000 users see the webpage containing the ad, then the company will owe $150, or the CPM, to the sponsoring website. The CPC fee structure typically converts to higher direct sales because only interested potential customers click, while the CPM fee structure ensures that a large number of potential customers see the ad, which is good for increasing brand or product awareness.

Assume that LEGO wants to place an ad on Rebrickable.com to promote a new LEGO product. Because LEGO is focused on generating awareness of the new product, it wants a CPM pricing structure based on the number of times that the ad is viewed. In 2020, Rebrickable.com had more than 80 million page views.[23] LEGO plans on spending $10,000

[22]Optimization and sentiment analysis typically involve advanced techniques. Therefore, they are discussed in Chapter 11.

[23]Source: Nathan, "Rebrickable 2020 Stats—Year in Review," January 1, 2021. https://rebrickable.com/blog/326/rebrickable-2020-stats-year-in-review/. Accessed 9/25/2021.

on the ads and wants at least 5 million views of the ad. Based on these parameters, what CPM rate should LEGO pay?

To answer this question, we will use Goal Seek in Excel **(Data > Forecast > What-if Analysis > Goal Seek)**, which works backwards to determine the item of interest; here, the CPM rate. Exhibit 7.21 presents the base spreadsheet (on the left) as well as the contents of the Goal Seek dialog box (on the right). Cell B2 sets the Marketing Budget to $10,000. The value of $2.50 acts as a placeholder in cell B3 until Goal Seek solves the CPM. The cost per 1 view (cell B4) is the CPM divided by 1,000. Total Views (cell B5) equals the Marketing Budget (cell B2) divided by the Cost per 1 view (cell B4). In the Goal Seek dialog box, we know that LEGO wants Total Views to equal 5 million, so $B$5 is the Set cell, and the To value is the 5,000,000 goal. Because we want to determine the CPM rate, we specify that variable in the By changing cell ($B$3).[24]

Exhibit 7.21
Goal-Seek Setup in Excel to Determine CPM

	A	B
1		
2	Marketing Budget	$ 10,000
3	Cost per 1,000 viewers (CPM)	$ 2.50
4	Cost per 1 view	=$B$3/1000
5	Total Views	=$B$2/$B$4

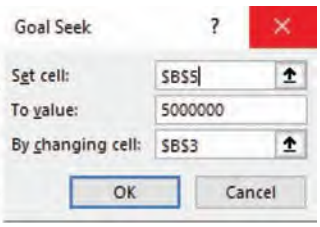


Exhibit 7.22 shows the answer generated by Goal Seek: a CPM rate of $2.00. Note that this amount is the maximum that LEGO should pay. This is a simple example; Goal Seek can be used to help solve much more complicated problems.

Exhibit 7.22
Goal-Seek Answer in Excel to Determine CPM

	A	B
1		
2	Marketing Budget	10,000
3	Cost per 1,000 viewers (CPM)	$ 2.00
4	Cost per 1 view	0.002
5	Total Views	5,000,000

MINI CASE: American Airlines' Sky-High Cash Flow Problem

In the 1980s, **American Airlines** needed to raise money quickly to survive. Its solution was the AAirpass membership program. The original 1981 AAirpass allowed members unlimited first-class travel for the rest of their lives for a one-time payment of $250,000 (equal to $715,000 today), with the option to purchase a companion pass for $150,000. Over the decades, American Airlines has sold this AAirpass membership in a variety of formats (it is now called Airpass), with later versions of the program allowing the same unlimited miles of travel but at a higher price tag ($3 million in 2004). Members, including

[24]The use of $ in an Excel formula makes the cell reference absolute so that it does not change.

billionaire **Shark Tank** investor and **Dallas Mavericks** owner Mark Cuban, have called the AAirpass one of the best purchases they ever made.[25]

While AAirpass may have been good for its members, an analysis of members' behavior shows that it has not been good for American Airlines. For example, two members, Steven Rothstein and Jacques Vroom, used their passes the most, flying almost 50 million miles over 25 years. Each flew hundreds of times a year, all over the world, sometimes just to have lunch with a friend in a foreign city such as London or Paris. Another member flew to London 16 times in one month! While the AAirpass provided the airlines with an immediate cash infusion, American Airlines lost millions of dollars during the AAirpass era, leading to changes in the program.[26] The current program, Airpass, is an annual membership (starting at $10,000) that provides premium seating, airport Admiral Club membership, on-demand booking, and no change or cancellation fees.[27]

Decision-Making with Data

How could American Airlines have used prescriptive analytics to evaluate the long-term profitability of AAirpass? How could prescriptive analytics have helped American Airlines determine how to structure a membership program that would be more profitable for the airline?

LAB CONNECTIONS

LABS 7.7 and **7.8** provide step-by-step Excel instructions for performing goal-seek analysis to help determine a specified variable based on assumptions.

- **LAB 7.7** determines an acceptable advertising rate.
- **LAB 7.8** determines the appropriate product price.

PROGRESS CHECK

17. What is the difference between CPC and CPM?
18. The LEGO analysis in this section assumes that LEGO wanted to promote awareness of a new product, making determining the CPM rate an appropriate analysis. What additional objectives may LEGO have with new products, and which technique(s) listed in Exhibit 7.20 would be appropriate to use with each objective?

[25]Source: Zachary Crockett, "The Rise and Demise of the AAirpass, American Airlines' $250k Lifetime Ticket," April 7, 2018, *The Hustle,* https://thehustle.co/aairpass-american-airlines-250k-lifetime-ticket/. Accessed 8/23/2021.

[26]This section is based on the following sources: Chris Loh, "The American Airlines AAirpass—The Unlimited Flight Pass That Didn't Work," *Simple Flying,* April 19, 2020, https://simpleflying.com/american-airlines-aairpass/. Accessed 8/23/2021. Zachary Crockett, "The Rise and Demise of the AAirpass, American Airlines' $250k Lifetime Ticket," April 7, 2018, *The Hustle,* https://thehustle.co/aairpass-american-airlines-250k-lifetime-ticket/. Accessed 8/23/2021.

[27]Source: https://airpass.aa.com/.

LO 7.8

Describe how marketing analytics reports results.

DATA VISUALIZATION

7.8 REPORT THE RESULTS

As noted in Chapter 6, there are numerous ways to report the results of data analysis. This section focuses on marketing dashboards. A **marketing dashboard** is a visual management tool that helps track and monitor KPIs to improve marketing effectiveness. Each company needs to decide not only which marketing KPIs are important for success but also the best way to present them on the dashboard.

One example of a marketing dashboard is **Google Trends** (https://trends.google.com/trends/?geo=US), which shows the number of times that a word is searched for using Google. Marketers use this information to direct traffic to the company's website by incorporating/indexing popular words into the company's website content, or by creating a marketing campaign using popular words. Exhibit 7.23 presents the dashboard for the Google search term "LEGO." The top graph shows how frequently the term "LEGO" has been entered in Google since 2004. The bottom visual shows the location of the users entering the term. For example, users in Utah search for "LEGO" more than any users in any other U.S. state or subregion. Google Trends allows the user to customize the time period and country as well as state, metro, or city location. Users can also identify another search term to "Compare," producing a side-by-side comparison in the graph.

Exhibit 7.23
Google Trends for "LEGO" Search Term in the United States

(Source: https://trends.google.com/trends/explore?date=all&geo=US&q=LEGO, Accessed 10/24/2021.)

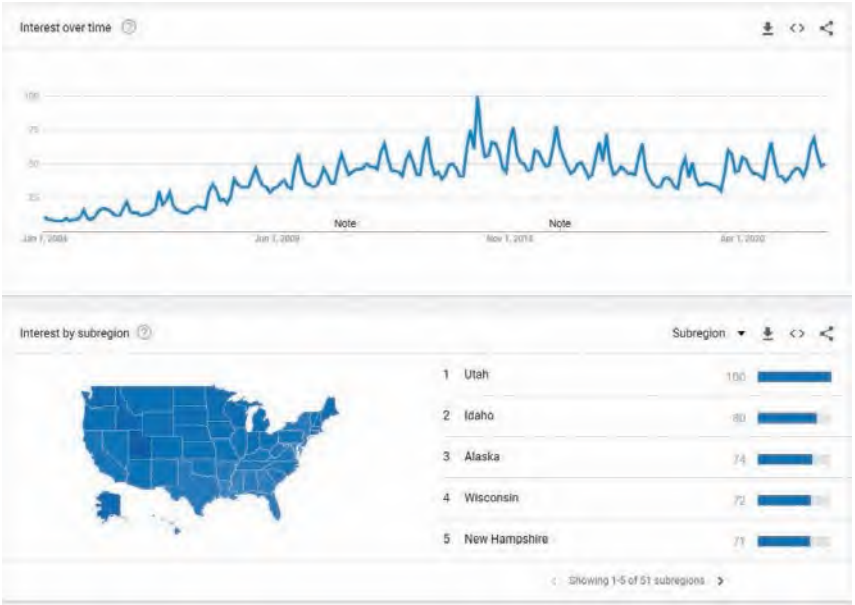


Exhibit 7.24 shows an example of a marketing dashboard used by **Klipfolio** to monitor the effectiveness of its digital marketing initiatives. This dashboard focuses on digital analytics, including a variety of KPIs on leads as well as Google Analytics on website usage and traffic.



Exhibit 7.24
An Example of a Marketing Dashboard for Klipfolio

(Source: https://www.klipfolio.com/resources/dashboard-examples. Accessed 9/25/2021.)

PROGRESS CHECK

19. What is the difference between a data visualization and a dashboard?
20. What KPIs could LEGO include in its marketing dashboard?

Chapter Summary

Marketing analytics is used to answer marketing questions. We provided examples of the types of questions that can be addressed with both internal data and external data across the four Ps of the marketing mix: product, price, place, and promotion. We also discussed the data available to address marketing questions and described the different types of analytics that can be performed, including descriptive, diagnostic, predictive, and prescriptive analytics. We provided examples of several different types of data visualizations that are useful for marketing, and we concluded with an example of reporting results using a marketing dashboard.

Key Terms

algorithm (462) A set of instructions on how to combine data.

bounce rate (456) The percentage of website visitors who enter a company's website and immediately leave.

brand (439) A phrase, symbol, logo, design, or word that uniquely identifies a company.

cluster analysis (458) A statistical technique that organizes data into clusters, or groups, in such a manner that all of the data within any cluster are similar, while data across different clusters are different.

conversion rate (444) The percentage of potential customers who actually make a purchase.

cost-per-click (CPC) fees (465) Internet advertising fees based on the number of clicks.

cost-per-mille (CPM) fees (465) Internet advertising fees based on the number of thousands of impressions or views.

customer journey (449) A visual representation of the experience that a customer has with a company, showing all of a customer's interactions or touchpoints with a company.

customer journey analytics (449) A type of marketing analytics that tracks and analyzes how customers used different channels to interact with a company across time.

customer relationship management (CRM) system (444) An information system for managing all interactions with past, current, and potential customers. It stores contact information, records sales and service issues, helps manage marketing campaigns to identify potential sales opportunities, and manages/records communications with customers across marketing channels.

decision tree (463) A decision-support tool that visually shows various outcomes from different (sequential) decisions.

drill(ing) down (457) A data-summarization technique that provides more views/breakdowns of the data. The data are broken down into more details when you drill down.

drill(ing) up (457) A data-summarization technique that provides fewer views/breakdowns of the data. The data are more summarized when you drill up.

drill(ing) through (457) A data-summarization technique that provides alternative views of the data.

Ishikawa diagram (fishbone, herringbone, or cause-and-effect diagram) (457) A diagram that captures the underlying factors (and the relationships) that may be the root cause of the problem. It should include all possible causes, including unrelated, correlated, contributing, and root causes.

key performance indicators (KPIs) (454) Numerical metrics used to gauge company performance.

marketing (440) The activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large.

marketing analytics (441) The processes and technologies that enable marketers to evaluate the success of their marketing initiatives by measuring performance using important business metrics.

marketing channels (442) Final product distribution points, such as wholesalers, online stores, or bricks-and-mortar retail stores.

marketing dashboard (468) A visual management tool that helps track and monitor KPIs (key performance indicators) to improve marketing effectiveness.

marketing mix (440) The key choices that a company makes to bring a product/service to market. It consists of product, price, place, and promotion (the four Ps).

place (440) Where a company sells a product and how the product is delivered to the market.

price (440) The amount that a customer pays for a product.

product (440) The good or service that a company offers to customers.

product life cycle (440) The four phases (introduction, growth, maturity, and decline) that products go through, beginning with development and ending with decline/market removal.

promotion (441) All of the marketing communication strategies and techniques used to convince customers that they need the product at the specified price. It includes advertising, public relations, sales promotions, special offers, and social media.

root cause analysis (456) A process used to understand what factor, or factors, caused a change.

sales funnel chart (460) An inverted pyramid chart used to show the steps in the sales process.

search engine optimization (SEO) (444) Identifying popular keywords, terms, and phrases that direct search engines to the company's website.

social media analytics (SMA) (447) The process of harvesting and analyzing user data from social media.

web analytics (444) The processes and technologies that extract and categorize quantitative and

ANSWERS TO PROGRESS CHECKS

1. Product, price, place, and promotion are the four components of the marketing mix. Product is the good/service being offered to customers. Price is the amount a customer pays for the product. Place refers to where a company sells a product and how the product will be delivered to the market. Promotion refers to all of the marketing communication strategies and techniques used to convince customers that they need the product at the specified price.
2. The type of product affects the price that can be charged, the place(s) the product can be sold, and the types of promotions that can be used. Price also affects the places where the product is sold at as well as the types of promotion used. Place affects the promotions used, because promotions maybe tailored to the place.
3. Potential ways **LEGO** can improve its marketing mix include developing products that mirror current trends or assist with building huge LEGO displays. Two examples would be a LEGO drone kit that actually flies or blueprints for an enormous build of the Empire State Building.
4. Potential marketing questions addressable with predictive analytics include "What are the expected unit sales in the different phases of the product life cycle?" and "What are forecasted unit sales based on season and location?"
5. An argument could be made to support any of the four Ps as being the most important for a company's success. For example, the product could be considered the most important because it is difficult to convince customers to buy a product they don't want or like.
6. Sources of internal data that can be used for marketing analytics include the company's ERP/ES, CRM system, company websites, and internally generated customer survey-based data about customer sentiment and customer satisfaction.
7. Sources of external data that can be used for marketing analytics include social media (such as **Facebook** and **YouTube**) as well as organizations that provide economic, opinion, and/or industry data. These include the U.S. Census Bureau, U.S. Bureau of Economic Analysis, U.S. Small Business Administration, United Nations, World Bank, Pew Research Center, **LexisNexis**, **Nielsen**, **Statista**, and **Experian**. For more sources, see https://courses.lumenlearning.com/marketing-spring2016/chapter/reading-marketing-data-sources/.
8. Companies can use data warehouses, data lakes, OLAP cubes, and customer journey analytics to store the data they will use to perform marketing analytics.
9. As indicated by the frown emoji, the following items need to be changed to improve the customer experience: contact travel department, book tickets, receive tickets, board plane, safety procedures, and takeoff. The following items need more data: book tickets, check-in at Paddington train station, expedited security and passport checks for frequent flyers, seating—wide comfy seats, disembark—reserve seat for return flight, passport & immigration fast track, customs fast track, driver checks into hotel for me, and hotel reconfirm flights back. A variety of data could be collected, including (security, passport check, disembark, customs, hotel check-in) elapsed times, customer satisfaction with the seats, and customer satisfaction with flight reconfirmation.
10. Many potential data sources might be important to **LEGO**. For example, social media analytics is extremely important to LEGO given its huge online experience and web community. Analyzing online comments and suggestions should help LEGO understand its customers' likes, dislikes, other sentiments, and desired products.
11. The bounce rate is the percentage of potential customers who visit a website and immediately leave without going to a second page. Conversion rate is the percentage of potential customers who ultimately make a purchase. High bounce rates negatively impact conversion rates. Marketers can reduce the bounce rate by improving a variety of items, including the website content, web page loading time, and links to website content.

12. **LEGO** may track many KPIs. According to the company's financial highlights for the first half of 2021, LEGO is focusing on developing a digitally enabled brand. Therefore, KPIs about the digital experience will be very important in monitoring the company's progress to that goal. Potential KPIs include website traffic, time spent on company website and individual web pages, number of web pages viewed per visit, number of return visitors versus new visitors, click rate on webpages (that is, how often viewers click to learn more), number of times a video is viewed, and keyword rankings.
13. All of the terms focus on how data are presented. Drilling down allows you to see a number broken down into smaller parts (moving from more general to more specific). Drilling down should provide a deeper understanding of a summary total. Showing total sales revenue by region, state, and store is an example of drilling down. Drilling up means summarizing data at a higher level (moving from more specific to more general). Drilling up should help provide an overall perspective on the data. Moving from customer-specific data (such as sales dollars, products purchased, or promotions used) to company totals for all customers is an example of drilling up. Drilling through allows you to see the same data presented in a different way and should help provide a better understanding of the data. If a company wants to look at data in a different manner than currently presented but at the same summarization level, drilling through is the appropriate method because it will allow the company to see the data in a different type of visual or report. Looking at a customer data set in both a chart and a heat map format is an example of drilling through.
14. LEGO could use diagnostic analytics in many ways to help determine why website traffic decreased. For example, LEGO could determine what keywords drive traffic to its website. It could use hypothesis testing to see if the traffic driven by a keyword has changed over time.
15. Of the techniques presented in this chapter, regression is the best method to determine how customer satisfaction (and changes in it) impact sales revenue.
16. Additional predictive analytics questions include: Is a customer about to churn (that is, stop buying)? Which advertisement will have the best impact on sales revenue? How is weather related to sales revenue?
17. CPC and CPM are both advertising rates for placing ads on search engines or websites. CPC stands for cost-per-click; the search engine or website charges a fee to the advertiser only when someone clicks on the ad. CPM stands for cost-per-mille; the search engine or website charges a fee anytime that the webpage is viewed (with the rate based on a fee per 1,000 views).
18. LEGO could have many different objectives and use many different predictive techniques. For example, if LEGO wanted to increase sales from ads on other websites, then determining the optimal CPC rate with Goal Seek would be appropriate.
19. A data visualization is one graphical representation of data, while a dashboard combines many data visualizations.
20. LEGO could include many potential KPIs in its marketing dashboard. The KPIs will change as LEGO's strategy changes. For example, LEGO is focusing on developing a digitally enabled brand. Potential KPIs include website traffic, time spent on company website and individual web pages, number of web pages viewed per visit, number of return visitors versus new visitors, click rate on web pages (that is, how often viewers click to learn more), number of times a video is viewed, and keyword rankings.

connect

Multiple-Choice Questions

1. (LO 7.1) Which phrase refers to the key choices that a company makes to bring a product/service to market?
   a. Marketing channels
   b. Marketing mix
   c. Product life cycle
   d. Root cause
2. (LO 7.8) What combines multiple data visualizations and other analyses to track multiple key performance indicators?
   a. Marketing dashboard
   b. Ishikawa diagram
   c. Marketing mix
   d. Scenario analysis
3. (LO 7.2) Which term refers to the final product distribution points?
   a. Conversion points
   b. Leads
   c. Marketing channels
   d. Marketing mix
4. (LO 7.2) Which type of analytics addresses the question of how a promotion will affect each marketing channel?
   a. Descriptive
   b. Diagnostic
   c. Predictive
   d. Prescriptive
5. (LO 7.3) A CRM system manages all of a company's interactions with which of the following?
   a. Chief executive officers
   b. Clerical staff
   c. The corporate suite (C-suite)
   d. Customers
6. (LO 7.3) Which of the following captures all interactions with a customer?
   a. Customer journey analytics
   b. Digital analytics
   c. Social media analytics
   d. Web analytics
7. (LO 7.4, 7.5) A graph displaying KPIs for several years by marketing channel is an example of which type of analytics?
   a. Descriptive
   b. Diagnostic
   c. Predictive
   d. Prescriptive

8. (LO 7.5) Which technique is used to conduct a root cause analysis?
   a. Decision tree
   b. Ishikawa diagram
   c. Scenario analysis
   d. Summarization
9. (LO 7.6) Which type of technique divides leads into those who are likely to purchase and those who are not?
   a. Classification
   b. Decision tree
   c. Regression
   d. Sensitivity analysis
10. (LO 7.7) Which type of analytics determines what input is needed to achieve a desired goal?
   a. Descriptive
   b. Diagnostic
   c. Predictive
   d. Prescriptive
11. (LO 7.6) To estimate units sold, a company analyzed its **Facebook** advertising expenditures and the number of units sold. The result of the analysis was the following equation:

$$\text{Units Sold} = 12 + 0.21\ \text{FacebookAdExpenditures}$$

   This is an example of which type of marketing analytics?
   a. Descriptive
   b. Diagnostic
   c. Predictive
   d. Prescriptive
12. (LO 7.6) To estimate units sold, a company analyzed its **Facebook** advertising expenditures and the number of units sold. The result of the analysis was the following equation:

$$\text{Units Sold} = 12 + 0.21\ \text{FacebookAdExpenditures}$$

   Which of the following statements is NOT true based on the equation?
   a. The company expects to sell 12 units if it spends zero dollars on Facebook ads.
   b. The company expects to sell 12.21 units if it spends zero dollars on Facebook ads.
   c. The company expects to sell 33 units if it spends 100 dollars on Facebook ads.
   d. The company expects to sell 222 units if it spends 1,000 dollars on Facebook ads.
13. (LO 7.6) To estimate units sold, a company analyzed its **Facebook** advertising expenditures and the number of units sold. The result of the analysis was the following equation:

$$\text{Units Sold} = 12 + 0.21\ \text{FacebookAdExpenditures}$$

   The company would also like to simultaneously analyze its **YouTube** advertising expenditures and the impact on units sold. Which of the following statements is true?
   a. The company can still use the same equation to estimate units sold.
   b. It is impossible to simultaneously analyze the impact of Facebook and YouTube advertising expenditures.
   c. A new analysis and equation are needed for just YouTube advertising expenditures.
   d. A new analysis and equation are needed that include both Facebook and YouTube advertising expenditures as independent variables.

14. (LO 7.5) A company calculated the correlation between sending paper catalogs to customers and sales revenue. The correlation coefficient was 0.92. What does this value indicate?
    a. Sending paper catalogs and sales revenue are weakly related.
    b. Sending paper catalogs is negatively related to sales revenue.
    c. Sending paper catalogs is positively related to sales revenue.
    d. There is no relationship between sending paper catalogs and sales revenue.
15. (LO7.7) A company calculated the net present value (NPV) of an advertising campaign as – $56,700. What decision might this NPV lead to?
    a. The advertising campaign should be implemented.
    b. The advertising campaign should not be implemented.
    c. The advertising campaign will increase sales.
    d. The advertising campaign will not affect sales.

Discussion Questions

1. (LO 7.1, 7.2) Describe the role of marketing and marketing analytics in a company.
2. (LO 7.3) A customer relationship management (CRM) system can be set up to include information about a (potential) customer from the person's public social media activity, including likes/dislikes and posts about the company and its competitors. Do you believe it is ethical for a company to collect this information? Why or why not?
3. (LO 7.4, 7.5) How does KPI usage differ between descriptive analytics and diagnostic analytics? Provide an example.
4. (LO 7.3) The European Union (EU) General Data Protection Regulation requires internet users to explicitly "opt in" to share their personal information with other parties, such as advertisers. In the United States, "opt out" is more likely to be used, meaning that personal information is automatically shared unless the user takes additional steps to curtail or prevent that sharing. Which policy do you think is better? Why? How does the choice affect companies?
5. (LO 7.3, 7.4, 7.5, 7.6) Assume that a company purchases a month's worth of ad time on a digital billboard in a major city. How could the company evaluate the campaign's impact on customers?
6. (LO 7.4, 7.5) What is the difference between a website bounce rate and cart abandonment rate? How are the terms related to conversion rate?
7. (LO 7.3, 7.4, 7.5, 7.6, 7.7, 7.8) IoT, which stands for Internet of Things, refers to a network of physical objects that have a unique ID, connect to the internet, and can gather and share information. Examples include **Fitbit**, **Smart Lock**, **Ring**, and **Roomba**. These devices offer great features, including remote operation via an owner's phone (or any other connected device). Think about the information collected by these devices. How can marketers potentially use the information gathered from an IoT?
8. (LO 7.3, 7.4, 7.5, 7.6, 7.7, 7.8) Have you ever noticed that almost immediately after you search for something on the internet, you get a pop-up ad for a related item? Why do you think marketers use pop-up ads?
9. (LO 7.4, 7.5) How does the use of pivot tables differ between descriptive analytics and diagnostic analytics? Provide an example.
10. (LO 7.3) Exhibit 7.4 presents an experience wheel to document the customer journey. What other visual methods can companies use to document the customer journey?

connect

Brief Exercises

1. (LO 7.1) Match each marketing question to the relevant marketing mix component (product, price, place, or promotion).

Marketing Question	**Component of the Marketing Mix** (Product, Price, Place, or Promotion)
What marketing channels should we use?	
Should we use a cost-plus, competitor-based, or premium strategy for determining what customers pay for our product?	
How should our advertisements represent the brand?	
What distribution channels need to be in place?	
In which stage of the life cycle is the product?	
In which cities should we have a store?	
What is the value of our product?	

2. (LO 7.4, 7.5, 7.6, 7.7) Match each marketing question to the analytics type (descriptive, diagnostic, predictive, or prescriptive) that can help answer it.

Marketing Question	**Analytics Type** (Descriptive, Diagnostic, Predictive, or Prescriptive)
Why did unit sales decrease in the Great Lakes region?	
What will our product sales be if we increase incentives for our sales force?	
What was the average sales per salesperson last year?	
What is the optimal sales incentive to maximize product sales?	
On average, how many units are sold per day?	
Why did the website conversion rate fall?	
What is the best way to increase website conversion rate?	

3. (LO 7.3) Match the marketing data to the marketing mix component (product, price, place, or promotion) that it would be most helpful for analyzing.

Marketing Data	**Component of the Marketing Mix** (Product, Price, Place, or Promotion)
Competitor price	
Order fulfillment time	
Google Trends for keywords describing company products	
Competitor ad spending	
Distribution cost per product	
Pew Research Center surveys	
U.S. Bureau of Economic Analysis data	

4. (LO 7.4, 7.5, 7.6, 7.7) Match the analytics type (descriptive, diagnostic, predictive, or prescriptive) to the analytics technique used.

Analytics Technique Used	**Analytics Type** (Descriptive, Diagnostic, Predictive, or Prescriptive)
Summarization of last year's sales data	
Pivot table for drill down and drill up of sales by product	
Sentiment analysis to see what customers think of a new product	
Hypothesis test to see if conversion rates have significantly changed from last period	
Root cause analysis of why product sales have fallen	
Regression analysis with the conversion rate as the dependent variable and website traffic as the independent variable	
Pivot table to summarize sales by region	

5. (LO 7.4, 7.5, 7.6, 7.7) Match the analytics type (descriptive, diagnostic, predictive, or prescriptive) to the analytics technique used.

Analytics Technique Used	**Analytics Type** (Descriptive, Diagnostic, Predictive, or Prescriptive)
Goal-seek analysis for CPC rate	
Ishikawa diagram for low market growth	
Decision tree for product development	
Average units sold	
Correlation between advertising and sales revenue	
Classification of new customers	
Histogram of the number of different products sold during the quarter	

6. (LO 7.4, 7.5, 7.6, 7.7) Match the Excel menu option (Data Analysis, What-if Analysis) to the analytics technique used.

Analytics Technique Used	**Excel Menu Option** (Data Analysis, What-if Analysis)
Average of sales for the last five years	
Goal-seek analysis to determine the desired CPM rate	
Scenario analysis to select the best advertising campaign	
Correlation of user clicks of internet advertisements and sales revenue	
Regression of the relationship between advertising expense and sales revenue	

Problems

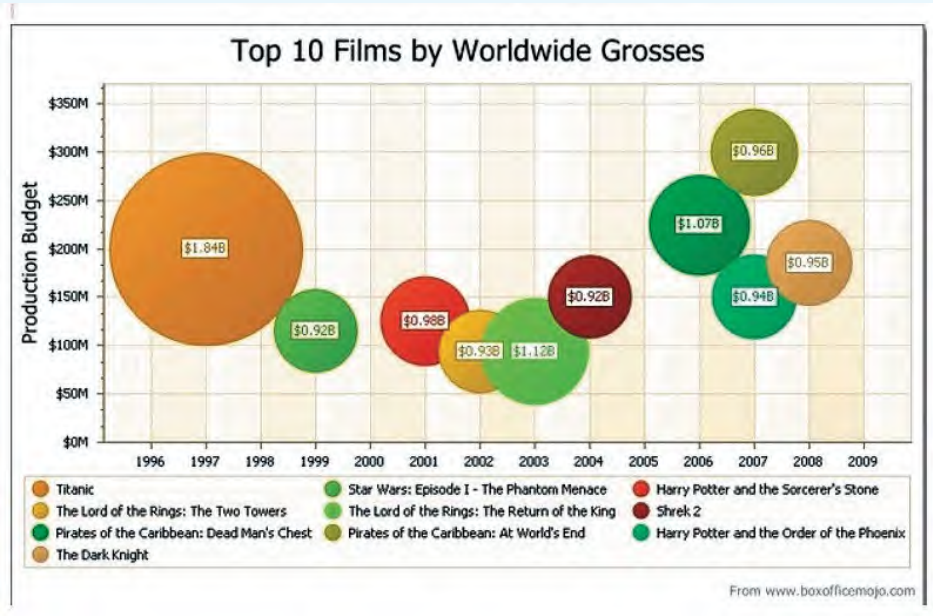
1. (LO 7.1, 7.4, 7.5, 7.6, 7.7) This chapter focuses on the traditional four Ps of the marketing mix (product, price, place, and promotion) and briefly introduced additional Ps: planning, process, people, and physical evidence. Process is the flow of activities that takes place when the company interacts with a customer. For example, the activities to return a product to **Home Depot** include: (1) Customer enters the Customer Service Desk line; (2) employee listens to customer request; (3) customer presents an item to return; (4) employee verifies the receipt and the item condition; (5) employee approves the return, obtaining manager help/approval when necessary; (6) employee enters information into the system including the reason for the return; and (7) employee provides the customer with a refund. Identify questions related to this process that Home Depot could ask that would require descriptive, diagnostic, predictive, and prescriptive analytics (one question for each type).
2. (LO 7.1, 7.4, 7.5, 7.6, 7.7) This chapter focuses on the traditional four Ps of the marketing mix (product, price, place, and promotion) and briefly introduced additional Ps: planning, process, people, and physical evidence. People refers to having the proper employees with appropriate skills sets or leadership skills. For example, the **Ford Motor Company** needs great salespeople to sell its cars and trucks. Identify questions related to the sales force that Ford could ask that would require descriptive, diagnostic, predictive, and prescriptive analytics to answer (one question for each type).
3. (LO 7.4, 7.5, 7.6, 7.7) **Panera Bread** sends out emails to customers to entice them to come and purchase food. Companies can tell if (and when) customers open emails. Assume that Panera Bread wants to use regression to determine if email open rates are related to subsequent food purchases. What are the independent and dependent variables in the regression? What additional variables would you suggest adding to the regression?
4. (LO 7.5, 7.6) This chapter describes sources of marketing data, including customers' web search history, social media posts, and purchase history. Analyzing this information can reveal private information, such as financial problems, illness, and pregnancy. The following story appeared in both *Fortune* and *The New York Times,* but it has never been confirmed by **Target**. The story provides a good opportunity to understand what could potentially be done with marketing data. A father found out that his teenage daughter was pregnant when Target began sending coupons for baby products to her (based on her purchases).[28] What type of diagnostic and predictive analytics could Target have used to determine that the teenager was pregnant? How can each of these techniques be used in marketing?
5. (LO 7.3) Exhibit 7.4 presents an example of a customer journey using **LEGO**'s experience wheel for a flight from London's Heathrow airport to New York City. For this problem, (1) select a company, brand, product, or experience and (2) document your own customer journey showing interactive points as well as your emotional response throughout the journey.
6. (LO 7.3) In this chapter, we discuss a customer journey, which is a visual representation of a customer's interactions with a company. A new process called UX Design is basically a customer journey taken to the n^{th} degree. The term "UX Design" was

[28]Ellen Rooney Martin, "The Ethics of Big Data," *Forbes,* March 27, 2014, https://www.forbes.com/sites/emc/2014/03/27/the-ethics-of-big-data/?sh=f1865226852e. Accessed 6/14/2021.

supposedly coined by Don Norman in the 1990s when he was VP of the Advanced Technology Group at Apple. UX refers to all aspects of a user's interaction with a company's services and products. It identifies how a person experiences everything when engaging with a product, including the person's attitudes and emotions. It requires seamless integration of engineering, marketing, interface design, graphical design, and industrial design. **Apple** used UX Design to make sure that Apple computer software was not only easy to use, but also that Apple computers easily fit into a person's life in terms of portability, size, visibility, sales experience, support experience, and other factors. UX Design builds on a customer journey to help optimize the customer's (total) experience at all touchpoints. Marketing analytics can help improve UX design by uncovering the underlying reasons why an experience is good or bad; it can also analyze the effectiveness of implemented solutions. One company that embraces UX design is **Disney**. Describe how Disney is a champion of UX Design in its theme parks. Consider a customer's journey as it begins from a Disney hotel through a day at one of its theme parks.

7. (LO 7.8) A well-designed data visualization can be very helpful in understanding a company's past performance so that managers can adjust current/future activities. The following data visualization was created to answer the following questions:[29]
   - How does the production budget required to make the film compare to worldwide grosses (revenues) for the film?
   - How do different blockbusters from 1996 to 2009 compare based on production budget and worldwide gross?

   How well does the data visualization answer these questions?

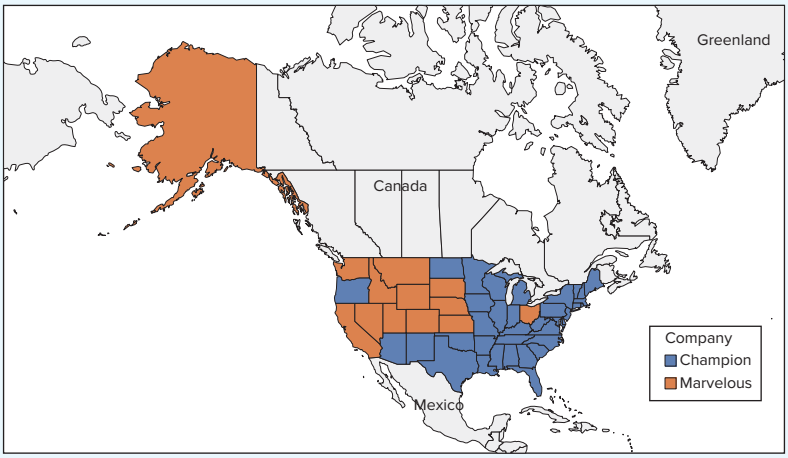


www.boxofficemojo.com.

[29]Source: https://thenextweb.com/news/7-most-common-data-visualization-mistakes. Accessed 11/02/2021.

8. (LO 7.8) Marketing data visualizations should help the company understand the impact of its marketing efforts. Examine the marketing data visualization below for two fictitious companies. Champion's marketing team created a visualization to show how Champion's market share compares to the market share of its main competitor, Marvelous. Explain how well this visualization answers the question: "Which company has the largest market share?" Do you have any recommendations to improve the data visualization?[30]

Market Leader by State



Use the following information to answer Problems 9 through 12:

Exhibits 7.6 and 7.7 analyze **LEGO** sets using data from **Rebrickable**. Exhibit 7.6 presents statistics for the number of LEGO sets released each year. Exhibit 7.7 shows the total number of sets released for a LEGO theme (for example, Star Wars). A subset of this data is provided for you in the LEGO set Data.xlsx file to answer Problems 9 through 12. An internal data model has been already defined. These data allow us to learn more about LEGO's product offerings. The data dictionary for the tables follows.[31]

Sets Table

set_num:	Unique identifier for a LEGO set
name:	Name of the LEGO set
year:	Year that the LEGO set was released
theme_id:	Unique identifier for LEGO theme
num_parts:	Number of unique LEGO parts used in the build

Themes Table

theme_id:	Unique identifier for LEGO theme
theme_name:	Name of LEGO theme (for example, Gear, Star Wars, City, Duplo)

9. (LO 7.4) Given the data dictionary above, identify the primary keys in each table and describe how the tables are connected. (If necessary, see Chapter 2 for a review of retrieving data from centralized databases.)
10. (LO 7.4) An important part of business analytics is understanding the data using descriptive analytics. Use Excel to answer the following questions about num_parts for LEGO sets released from 1949 to 2021. *Hint:* Make sure the data set includes only these years.
    a. How many observations are in the original data set in the "Sets" worksheet tab? How many observations are used to answer parts b and c of this problem? Why is the number of observations different?

[30]Adapted from https://thenextweb.com/news/7-most-common-data-visualization-mistakes. Accessed 11/02/2021.

[31]Note that the data from https://rebrickable.com/downloads/ have been simplified for the textbook.

b. Which set_num includes the largest number of unique parts (num_parts)?

c. What is the average number and median number of unique parts (num_parts) used in the sets?

11. (LO 7.4) Use Excel to answer the additional questions about num_parts for LEGO sets released from 1949 to 2021.

    a. What is the average number of parts (num_parts) used for each theme?

    b. Which theme has the largest average number of parts?

    c. What is the average number of parts included in the Star Wars themed sets? Note that four different Star Wars themed set categories should be included.

12. (LO 7.4) Use Tableau to create a bar chart that answers this question: How many LEGO sets with a Star Wars theme were released each year from 1949 to 2021? *Hint:* Four different Star Wars themed set categories should be included. What information does the bar chart reveal about Star Wars themed sets?

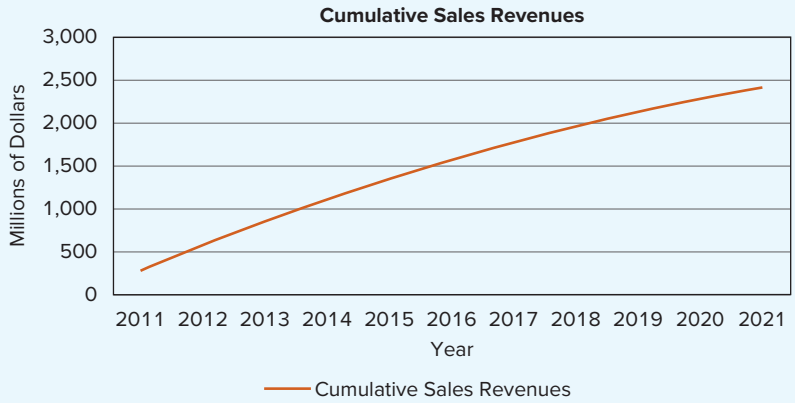
Data Visualization Exercise



DATA VISUALIZATION

1. When companies sell a product or service, they collect money from customers. Collectively, this money is called sales revenue. Changes in sales revenue over time are good indicators of how well a company is doing. Companies generally want annual sales revenue to increase over time, indicating that they are selling more products.

    a. What does the following visualization tell us about how well a company is doing in terms of annual sales revenue?



    b. Are annual sales revenues increasing or decreasing?

    c. Is there a better visualization to show how annual sales revenues are changing over time? Using Cumulative Sales data.xlsx, create a data visualization that shows the real pattern of annual sales revenue over time.

CHAPTER 7 LABS

LAB 7.1 **Excel:** Descriptive Analytics: Analyzing Company Historical Performance

LAB 7.2 **Excel:** Descriptive Analytics: Using a Pivot Table to Analyze Historical Performance by Product Size and Year

LAB 7.3 **Tableau:** Descriptive Analytics: Using a Histogram to Evaluate Process Time

LAB 7.3 **Power BI:** Descriptive Analytics: Using a Histogram to Evaluate Process Time

LAB 7.4 **Excel:** Diagnostic Analytics: Analyzing the Steps in the Sales Process with a Sales Funnel Chart

LAB 7.5 **Tableau:** Diagnostic Analytics: Examining Pricing Strategy with Cluster Analysis

LAB 7.6 **Excel:** Predictive Analytics: Predicting Sales Revenue from Advertising Expense

LAB 7.6 **Tableau:** Predictive Analytics: Predicting Sales Revenue from Advertising Expense

LAB 7.7 **Excel:** Prescriptive Analytics: Calculating Internet CPM Rate Using Goal Seek

LAB 7.8 **Excel:** Prescriptive Analytics: Calculating Product Price Using Goal Seek

Lab 7.1 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Descriptive Analytics: Analyzing Company Historical Performance

Keywords

Descriptive Analytics, Histogram, Excel Analysis ToolPak

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" In this lab, you will use the Excel Analysis ToolPak to generate descriptive analytics of the data described below.

Several years ago, Rob built a table for his son to use in building LEGO creations. The table allowed his son to stand while building, provided a frame to hold LEGO base plates, and got the LEGOs off the floor. Several of Rob's neighbors saw the table and asked him to build tables for their children. The tables grew in popularity and Rob eventually quit his job in 2018 and founded LeTable Inc. He now works full-time building custom LEGO tables.

Customers select from four different sizes of tables:

1. Small, which holds 4 base plates
2. Medium, which holds 8 base plates
3. Large, which holds 16 base plates
4. Deluxe, which has multiple levels as well as conduits for electric lights

Customers can also select from five different colors:

1. Blue
2. Red
3. Yellow
4. Natural (polyurethane only)
5. Stained

Rob would like an analysis of how the company has performed over the company's first five years (2018–2022).

Required

1. Calculate descriptive analytics for selling price, cost, and order process time.
2. Create a histogram for order process time.
3. Create a written report to present your results to Rob.

Specify the Question

What is the average, median, and mode selling price and cost for LeTable Inc. over the past five years? What is the standard deviation for selling price and cost? How long does it take LeTable Inc. to make a table after the order is placed (that is, process time)?

Process Time: The time it takes to fill an order. It is measured as the number of days between receiving the order and shipping the table to the customer.

Obtain the Data

This lab uses the data in Lab 7.1 Data.xlsx. The data dictionary for this file is below. Note that the order dates range from 1/5/2018 to 12/31/2022 for a total of 19,995 orders. In addition, some orders were not completed and shipped until 2023.

Data Dictionary

OrderID:	Sequential number identifying the sales order
CustomerName:	First and last name of the customer[32]
ZipCode:	5-digit zip code of the customer
OrderDate:	Date that the order was placed in MM/DD/YYYY format
ShipDate:	Date that the completed order was shipped in MM/DD/YYYY format
Color:	Color of the table (blue, red, yellow, natural, stained)
Size:	Table size (small, medium, large, deluxe)
SellingPrice:	Price the table was sold for
Cost:	Cost to make the table

Analyze the Data

Analysis Task 1: Install the Analysis ToolPak.

The first step is to install the Analysis ToolPak if it is not already installed. Textbook Appendix E provides instructions for installing the ToolPak.

Analysis Task 2: Calculate Process Time.

1. Before calculating descriptive statistics, add an additional column for process time—that is, the number of days it took to make the table after it was ordered. Type "ProcessTime" in cell J1. Type the formula "=E2-D2" in cell J2.

[32]For Labs 7.1, 7.2, 7.3, and 7.4, all names were generated using https://www.databasetestdata.com/.

2. Copy[33] the formula down column J. To copy down the entire column, double-click on the lower-right corner of cell J2. The formula should copy down the column to the last entry. Lab Exhibit 7.1.1 shows what the spreadsheet should look like.

Lab Exhibit 7.1.1

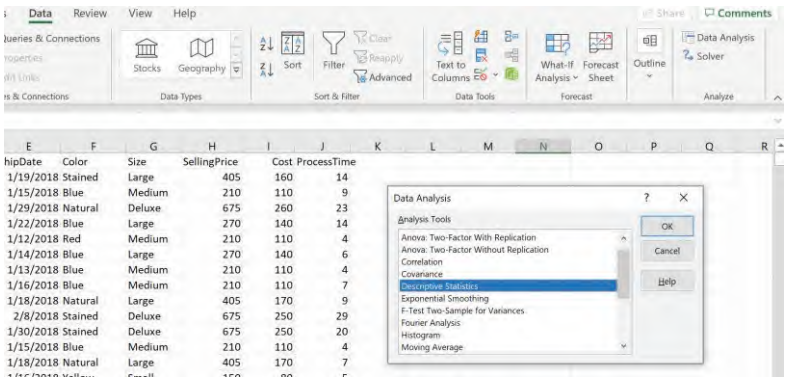
	A	B	C	D	E	F	G	H	I	J
1	OrderID	CustomerName	Zip	OrderDate	ShipDate	Color	Size	SellingPrice	Cost	ProcessTime
2	1	Mrs. Jammie Kunze	75000	1/5/2018	1/19/2018	Stained	Large	405	160	14
3	2	Nicolette Kassulke	75001	1/6/2018	1/15/2018	Blue	Medium	210	110	9
4	3	Dr. Theresia Wunsch	75002	1/6/2018	1/29/2018	Natural	Deluxe	675	260	23
5	4	Vesta Murazik	75006	1/8/2018	1/22/2018	Blue	Large	270	140	14
6	5	Madie Marvin	75009	1/8/2018	1/12/2018	Red	Medium	210	110	4
7	6	Brandon Gerhold	75019	1/8/2018	1/14/2018	Blue	Large	270	140	6
8	7	Louisa Windler	75020	1/9/2018	1/13/2018	Blue	Medium	210	110	4
9	8	Brown Hessel	75022	1/9/2018	1/16/2018	Blue	Medium	210	110	7
10	9	Telly Harmann	75023	1/9/2018	1/18/2018	Natural	Large	405	170	9
11	10	Gaylord Borer	75029	1/10/2018	2/8/2018	Stained	Deluxe	675	250	29
12	11	Miss Melvin Johns	75030	1/10/2018	1/30/2018	Stained	Deluxe	675	250	20
13	12	Jett Aufderhar	75032	1/11/2018	1/15/2018	Blue	Medium	210	110	4
14	13	Fabiola Rodriguez	75032	1/11/2018	1/18/2018	Natural	Large	405	170	7
15	14	Lauryn Spencer	75033	1/11/2018	1/16/2018	Yellow	Small	150	80	5
16	15	Ms. Jacques Lang	75040	1/12/2018	1/22/2018	Stained	Medium	315	130	10

Analysis Task 3: Calculate Descriptive Statistics Using the Excel Analysis ToolPak.

You are now ready to calculate descriptive statistics using the Analysis ToolPak.

1. In the menu, select **Data > Analyze > Data Analysis.**
2. Select Descriptive Statistics, as shown in Lab Exhibit 7.1.2. Click OK.

Lab Exhibit 7.1.2

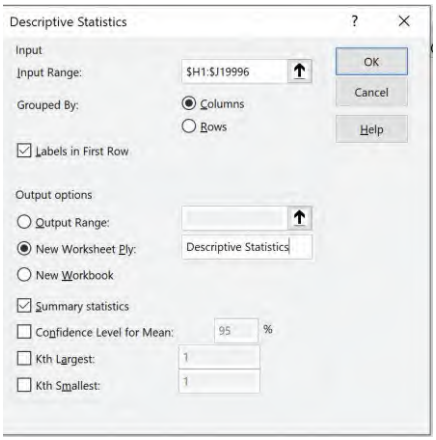


3. Descriptive Statistics work only on numeric fields, so only SellingPrice, Cost, and ProcessTime are appropriate fields. In the Descriptive Statistics dialog box, enter the Input Range as "$H$1:$J$19996". Check the box for Labels in the First Row because the column labels are in row 1. Select New Worksheet Ply: to have the results automatically placed in a new worksheet tab. (*Note:* To put the descriptive statistics on the same worksheet, select Output Range and identify the place you want to show the results. To have the results placed in a completely new workbook or file without the Data sheet, select New Workbook.) Type "Descriptive Statistics" in the field by New Worksheet Ply to automatically name the new worksheet. Check the box for Summary

[33]See this video: https://www.lynda.com/Excel-tutorials/Copy-data-formulas-down-column-instantly/431183/492958-4.html.

statistics to generate basic statistics, including mean, minimum, and maximum. The Descriptive Statistics dialog box should look like Lab Exhibit 7.1.3. Click OK.

Lab Exhibit 7.1.3



Excel should automatically generate the descriptive statistics. Lab Exhibit 7.1.4 shows the output for SellingPrice. Note the Average (Mean) SellingPrice over the last five years was $287.31 with a minimum of $150 (Minimum) and a maximum of $675 (Maximum). Thus, the Range is 675 − 150 = 525. The most common SellingPrice is $210 (Mode). Median represents the midpoint. Thus, half of the selling prices are above $270 and half are below $270.

Lab Exhibit 7.1.4

	A	B
1	*SellingPrice*	
2		
3	Mean	287.3121
4	Standard Error	0.807096
5	Median	270
6	Mode	210
7	Standard Deviation	114.1264
8	Sample Variance	13024.83
9	Kurtosis	2.98511
10	Skewness	1.57584
11	Range	525
12	Minimum	150
13	Maximum	675
14	Sum	5744805
15	Count	19995

Since 2018, LeTable Inc. has had 19,995 orders (Count) and a total of $5,744,805 (Sum) in sales. The remaining measures provide insights into the dispersion of the data, as follows:

A higher Standard Deviation indicates that the data are more dispersed. The standard deviation is the square root of the Sample Variance, which measures how far numbers are spread out from the average value. For this data set, the majority of selling prices are within one standard deviation of the mean. Specifically, the majority of the selling prices are between $401 ($287 + $114) and $173 ($287 − $114).

Kurtosis and Skewness describe the shape of the distribution, helping determine the likelihood of an event falling in a tail of a probability distribution. Kurtosis measures the thickness of the tails of the distribution. A kurtosis value of 2.98511 indicates that the

distribution is close to a normal distribution, which has a kurtosis value of 3. Skewness equals 1.57584, indicating that the distribution is positively (right) skewed. (A negative skewness value indicates that the distribution is skewed to the left.)

The Descriptive Statistics option automatically calculates the statistics for the entire column. To get a breakdown of a column, for example selling price by color, you will have to manually split up the column into ranges or use a different Excel functionality (such as a pivot table).


DATA VISUALIZATION

Analysis Task 4: Create a Histogram for ProcessTime Using the Excel Analysis ToolPak.

Now let's use a histogram to analyze how long it takes to complete an order. To create a histogram of ProcessTime, bins must be defined first. Bins are categories that group the data for purposes of analyzing. For ProcessTime, it would be helpful to group the number of processing days in bins of 5 days.

1. Return to the Data worksheet.
2. In cell M1, type "Bins". In cell M2 type "0", in cell M3 type "5", in cell M4 type "10", and continue until you reach 30 in cell M8. Lab Exhibit 7.1.5 shows what your spreadsheet should look like.

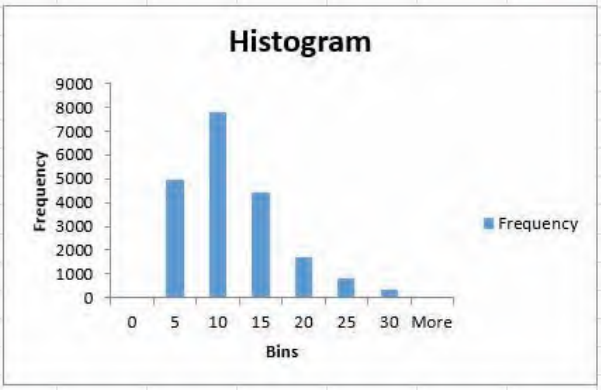
Lab Exhibit 7.1.5

M
Bins
0
5
10
15
20
25
30

3. In the menu, select **Data > Analyze > Data Analysis.**
4. Select Histogram. Click OK.
5. In the Histogram dialog box, in the Input Range field type "$J$1:$J$19996" (the ProcessTime column). In the Bin Range field enter "$M$1:$M$8". Check the Labels box because the label Bins is in cell M1. Select New Worksheet Ply and (optionally) type "Histogram" to automatically name the new worksheet. For this exercise, check Chart Output. Click OK. A histogram is automatically created in a new worksheet titled Histogram. The new worksheet shows the frequencies in a tabular format as well as a chart, as shown in Lab Exhibit 7.1.6. However, a careful inspection reveals that the graph is really a bar chart because there are gaps between the bars.

Lab Exhibit 7.1.6

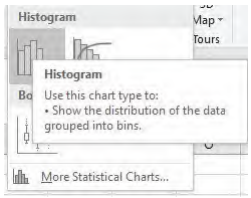
	A	B
1	*Bins*	*Frequency*
2	0	0
3	5	4929
4	10	7773
5	15	4409
6	20	1696
7	25	831
8	30	351
9	More	0



So, let's use an alternative method to create a real histogram.

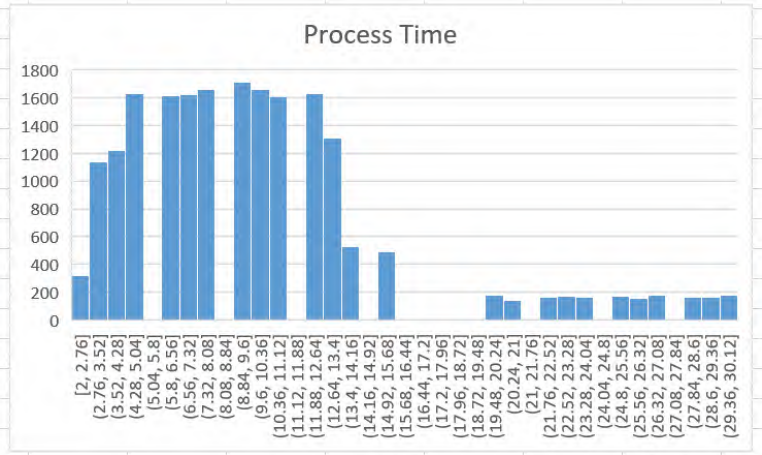
6. Return to the Data worksheet and highlight the entire Process column (J1:J19996). Before completing the next step, make sure you are at the top of the data, so the chart inserted is at the top of the worksheet.
7. From the menu, select **Insert > Charts > Statistic Chart.**
8. Select Histogram, as shown in Lab Exhibit 7.1.7.

Lab Exhibit 7.1.7



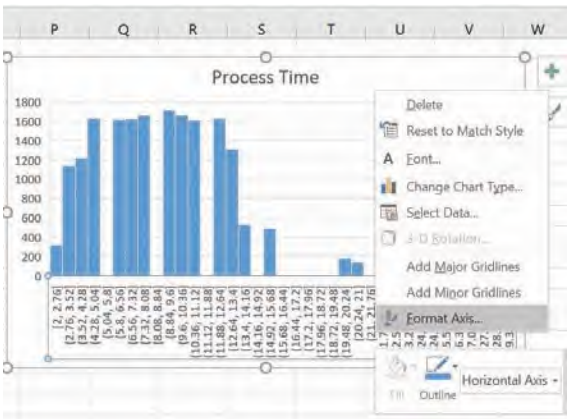
9. A histogram is created as shown in Lab Exhibit 7.1.8.

Lab Exhibit 7.1.8



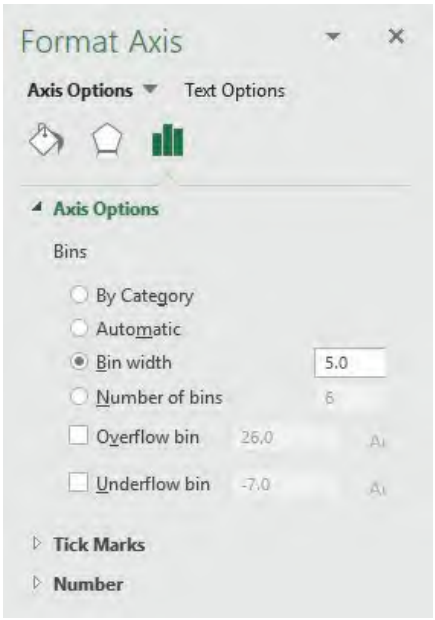
10. To change the title, right-click on the Chart Title, select Edit Text, and type "Process Time".
11. Notice that a very small bin size of 0.76 was used. Let's change the bin size. Right-click on the horizontal axis. A dialog box should appear. Select Format Axis, as shown in Lab Exhibit 7.1.9.

Lab Exhibit 7.1.9



12. The Format Axis dialog box appears. Select Bin width. Type "5". The Format Axis dialog box should look like Lab Exhibit 7.1.10. Press Enter.

Lab Exhibit 7.1.10



13. A real histogram appears! Now we need to add labels for our axes. Click on the chart box and a + should appear to the right. Click on the +. In the Chart Elements dialog box, check Axis Titles. Axis Title should appear along the vertical and horizontal axes. Click on the horizontal Axis Title and type "Days". Click on the vertical Axis Title and type "Number of Orders".
14. You may also want to format the vertical axis to the format of X,XXX. To do so, right-click on the numbers in the vertical axis, then select Format Axis. In the Format Axis dialog box, select Number. Set Category to Number, Decimal places to 0, and check Use 1000 Separator (,). Click outside of the dialog box to close it.

Report the Results: Visualize the Histogram Using a Pivot Chart

You have just created descriptive statistics for selling price, cost, and process time for LeTable Inc. for the years 2018 to 2022. In addition, you created a histogram for process time. Rob can use this information to see (1) how the company has performed over the past five years and (2) the timeliness with which the company delivers tables to customers. To share your analysis with Rob, prepare a formal report in Word including screenshots of the descriptive statistics tables and histograms that you created. The report should explain the results of your analysis and address the questions that he asked. In addition, you should recommend additional analysis that you think would help Rob make decisions about his business.

Assessment

1. Create a written report to present your results to Rob. Name your formal report in Word "Lab 7.1 Excel Submission.docx".
2. Answer the questions in Connect and upload your Lab 7.1 Excel Submission.docx via Connect, if assigned.

Alternate Lab 7.1 Excel

Descriptive Analytics: Analyzing Company Historical Performance

Apply the same steps in Lab 7.1 to the Alt Lab 7.1 Data.xlsx data set.

Process Time: The time it takes to fill an order. It is measured as the number of days between receiving the order and shipping the table to the customer.

Specify the Question

What is the average, median, and mode selling price and cost for LeTable Inc. over the past five years? What is the standard deviation for selling price and cost? How long does it take LeTable Inc. to make a table after the order is placed (that is, process time)?

Required

1. Calculate descriptive analytics for selling price, cost, and order process time.
2. Create a histogram for order process time.
3. Create a written report to present your results to Rob.

Assessment

1. Create a written report to present your results to Rob. Name your formal report in Word "Alt Lab 7.1 Excel Submission.docx".
2. Answer the questions in Connect and upload your Alt Lab 7.1 Excel Submission.docx via Connect, if assigned.

Lab 7.2 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Descriptive Analytics: Using a Pivot Table to Analyze Historical Performance by Product Size and Year

Keywords

Descriptive Analytics, Pivot Table

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" Lab 7.1 used the Analysis ToolPak in Excel to generate descriptive analytics for marketing data. This lab uses Excel pivot tables to summarize the data at different levels, allowing drill down and drill up (also known as roll up).

Several years ago, Rob built a table for his son to use in building LEGO creations. The table allowed his son to stand while building, provided a frame to hold LEGO base plates, and got the LEGOs off the floor. Several of Rob's neighbors saw the table and asked him to build tables for their children. The tables grew in popularity and Rob eventually quit his job in 2018 and founded LeTable Inc. He now works full-time building custom LEGO tables.

Customers select from four different sizes of tables:

1. Small, which holds 4 base plates
2. Medium, which holds 8 base plates
3. Large, which holds 16 base plates
4. Deluxe, which has multiple levels as well as conduits for electric lights

Customers can also select from five different colors:

1. Blue
2. Red
3. Yellow
4. Natural (polyurethane only)
5. Stained

Rob would like to see an analysis of his first five years of operations (2018–2022).

Required

1. Create a pivot table for total sales revenue, average selling price, average process time, and total gross profit by table size, year, and quarter.
2. Create a pivot chart for total sales revenue and total gross profit by table size, year, and quarter.

Specify the Question

Total Gross Profit: Gross profit for a table is sales price minus the cost of making the table. Total gross profit sums all of the gross profits for tables sold.

What are total sales revenue, average selling price, average process time, and total gross profit by table size, year, and quarter?

Obtain the Data

To compare the descriptive analytics generated automatically by the Descriptive Analytics option of the Analysis ToolPak and Excel pivot tables, this lab uses the same data as in Lab 7.1. The replicated data are contained in Lab 7.2 Data.xlsx.

The data dictionary for this file is below. Note that the order dates range from 1/5/2018 to 12/31/2022 for a total of 19,995 orders. In addition, some orders were not completed and shipped until 2023.

Data Dictionary

OrderID:	Sequential number identifying the sales order
CustomerName:	First and last name of the customer[34]
ZipCode:	5-digit zip code of the customer
OrderDate:	Date that the order was placed in MM/DD/YYYY format
ShipDate:	Date that the completed order was shipped in MM/DD/YYYY format
Color:	Color of the table (blue, red, yellow, natural, stained)
Size:	Table size (small, medium, large, deluxe)
SellingPrice:	Price the table was sold for
Cost:	Cost to make the table

Analyze the Data

Analysis Task 1: Calculate Process Time.

We first need to compute the process time for each sale (the number of days it took Rob to make the table after it was ordered).

1. Go to cell J1 and type the title "ProcessTime".
2. In cell J2, we'll compute the process time on each line item sold. To do so, we must subtract the OrderDate in cell D2 from the ShipDate in cell E2. Insert the formula "=E2-D2" and press enter.
3. Copy the formula down column J.

Analysis Task 2: Calculate Gross Profit.

We also need to compute the gross profit for each sale (that is, how much profit is left after covering the costs of making a table).

1. Go to cell K1 and type the title "GrossProfit".
2. In cell K2, we'll compute the gross profit on each line item sold. To do so, we must subtract the Cost in cell I2 from the SellingPrice in cell H2. In cell K2, type the formula "=H2-I2" and press enter.
3. Copy the formula down column K.

Lab Exhibit 7.2.1 shows what the spreadsheet should look like.

Lab Exhibit 7.2.1

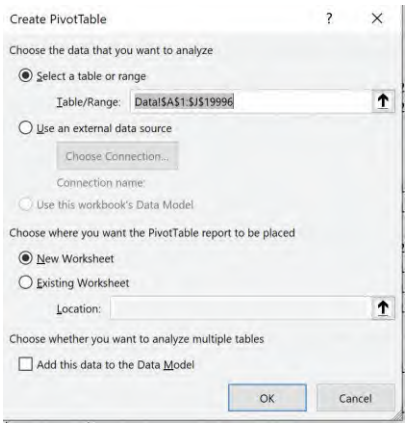
	A	B	C	D	E	F	G	H	I	J	K
1	OrderID	CustomerName	Zip	OrderDate	ShipDate	Color	Size	SellingPrice	Cost	ProcessTir	GrossProfit
2	1	Mrs. Jammie Kunze	75000	1/5/2018	1/19/2018	Stained	Large	405	160	14	245
3	2	Nicolette Kassulke	75001	1/6/2018	1/15/2018	Blue	Medium	210	110	9	100
4	3	Dr. Theresia Wunsch	75002	1/6/2018	1/29/2018	Natural	Deluxe	675	260	23	415
5	4	Vesta Murazik	75006	1/8/2018	1/22/2018	Blue	Large	270	140	14	130
6	5	Madie Marvin	75009	1/8/2018	1/12/2018	Red	Medium	210	110	4	100
7	6	Brandon Gerhold	75019	1/8/2018	1/14/2018	Blue	Large	270	140	6	130
8	7	Louisa Windler	75020	1/9/2018	1/13/2018	Blue	Medium	210	110	4	100
9	8	Brown Hessel	75022	1/9/2018	1/16/2018	Blue	Medium	210	110	7	100
10	9	Telly Harmann	75023	1/9/2018	1/18/2018	Natural	Large	405	170	9	235
11	10	Gaylord Borer	75029	1/10/2018	2/8/2018	Stained	Deluxe	675	250	29	425
12	11	Miss Melvin Johns	75030	1/10/2018	1/30/2018	Stained	Deluxe	675	250	20	425
13	12	Jett Aufderhar	75032	1/11/2018	1/15/2018	Blue	Medium	210	110	4	100
14	13	Fabiola Rodriguez	75032	1/11/2018	1/18/2018	Natural	Large	405	170	7	235
15	14	Lauryn Spencer	75033	1/11/2018	1/16/2018	Yellow	Small	150	80	5	70
16	15	Ms. Jacques Lang	75040	1/12/2018	1/22/2018	Stained	Medium	315	130	10	185

[34]For Labs 7.1, 7.2, 7.3, 7.4, and 7.5, all names were generated using https://www.databasetestdata.com/.

Analysis Task 3: Create the Pivot Table.

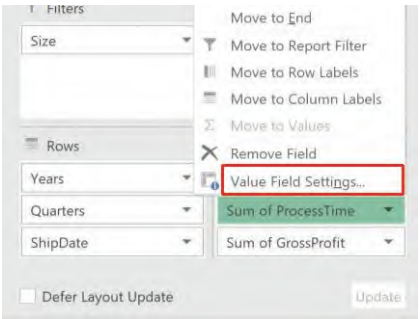
1. Click somewhere in the data.
2. From the menu, select **Insert > Tables > PivotTable.**
3. Keep the Table/Range that automatically filled in because you clicked in the data first. Click New Worksheet. The Create PivotTable dialog box should look like Lab Exhibit 7.2.2.

Lab Exhibit 7.2.2



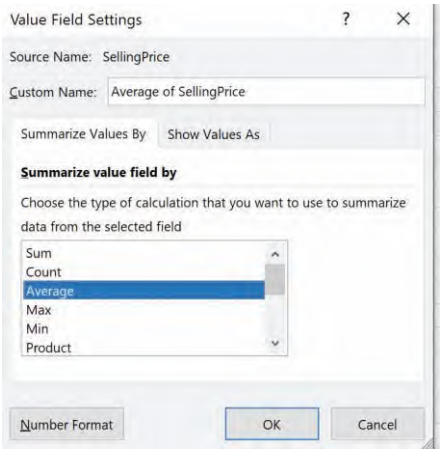
4. Click OK.
5. A new worksheet should appear with the PivotTable Fields dialog box on the right side. Drag ShipDate to Rows. Years and Quarters should automatically appear in Rows, allowing you to look at the data by year or by quarter. Note that some orders were shipped in 2023. Selecting ShipDate instead of OrderDate will allow you to see the ProcessTime for all 2022 orders.
6. Drag SellingPrice, ProcessTimeb, and GrossProfit to ΣValues. These are the fields that Excel will generate (selected) descriptive statistics for.
7. Drag Size to Filters, which allows the analysis to be done by table size.
8. Note that SellingPrice, ProcessTime, and GrossProfit defaulted to "Sum of" in ΣValues. Rob would like to see the average SellingPrice and average ProcessTime. To change the "Sum of" to "Average of", click on the right arrow next to the Sum of SellingPrice and select Value Field Settings, as shown in Lab Exhibit 7.2.3.

Lab Exhibit 7.2.3



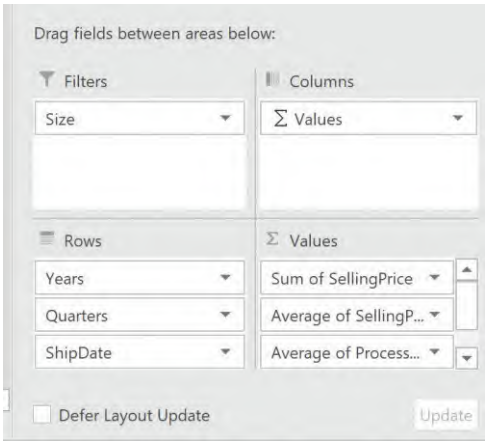
9. In the Value Fields Settings dialog box, change Sum to Average, as shown in Lab Exhibit 7.2.4.

Lab Exhibit 7.2.4



10. Click OK.
11. Repeat the process to change ProcessTime to "Average of".
12. Rob also wants to see total SellingPrice, so drag SellingPrice to the top of ΣValues. Sum of SellingPrice will give the total sales. Your Pivot Table Fields dialog box should look like Lab Exhibit 7.2.5. (If you scroll down in ΣValues, you will see Sum of Gross Profit.)

Lab Exhibit 7.2.5

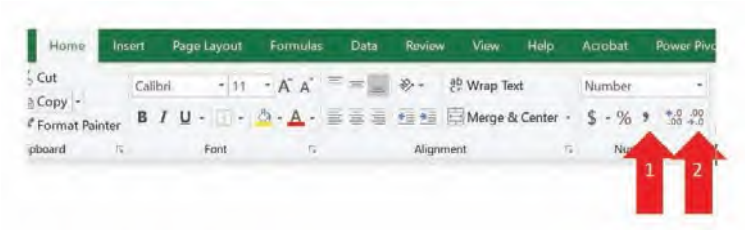


Analysis Task 4: Format the Pivot Table.

Follow these steps to improve the formatting of the table.

1. Format Sum of SellingPrice (column B) and Sum of GrossProfit (column E) with a "," and no decimal places. Select the columns you want to change and then use Lab Exhibit 7.2.6 to find the location of the "," and "Decrease Decimal" in the Home tab. Click the option "," (shown in Lab Exhibit 7.2.6 with red arrow 1) and "Decrease Decimal" (shown in Lab Exhibit 7.2.6 with red arrow 2) as many times as necessary to make the decimals disappear.

Lab Exhibit 7.2.6



2. Format Average of SellingPrice (column C) and Average of ProcessTime (column D) as two decimal places. Select the columns you want to change. Click the "Decrease Decimal" (shown with red arrow 2 in Lab Exhibit 7.2.6) as many times as necessary to make (only) two decimals show.

Analysis Task 5: Use the Pivot Table.

Unlike the Descriptive Analytics in the Analysis ToolPak, a pivot table allows you to (automatically) analyze the data at different levels. In other words, you can drill down or drill up (roll up) to the desired level.

1. The Size filter has a dropdown list that allows you to choose which product size you would like to analyze. If you click on the + next to a year, then the year drills down to the quarters.
2. If you click on the + by the quarter, then the quarter drills down to the months.
3. If you want to see the transactions making up a month's orders, click on the amount in the month row and the individual transactions for that month will appear in a new worksheet.

Lab Exhibit 7.2.7 shows data for the Deluxe table model with Quarter 1 of 2018 expanded.

Lab Exhibit 7.2.7

	A	B
1	Size	Deluxe
2		
3	Row Labels	Sum of SellingPrice Averag
4	2018	83,700
5	Qtr1	20,250
6	Jan	1,350
7	Feb	8,100
8	Mar	10,800
9	Qtr2	23,400
10	Qtr3	25,200
11	Qtr4	14,850
12	2019	112,050
13	2020	171,900
14	2021	284,850
15	2022	335,025
16	2023	22,500
17	Grand Total	1,010,025



DATA VISUALIZATION

Report the Results: Visualize Annual Sales Revenue and Gross Profit

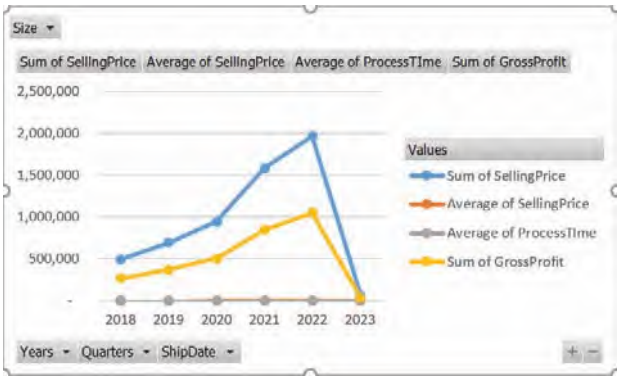
You have just created a pivot table for selling price, process time, and gross profit for LeTable Inc. Rob can use the pivot table to drill down and drill up to analyze historical sales. To help report the results to Rob, we're going to create a visualization of Total SellingPrice and GrossProfit using a pivot chart. Follow these steps:

1. Create a copy of the pivot table tab. Label it "Pivot Chart". Complete the following steps in this tab.

2. Select a cell in the pivot table.
3. From the menu, select **PivotTable > Tools Analyze > Tools > PivotChart.**
4. Select Line with Markers (the fourth chart).
5. Click OK.

The resulting pivot chart will be as shown in Lab Exhibit 7.2.8.

Lab Exhibit 7.2.8



6. Note that the averages and sums are shown in the same graph. Let's include only the sums in this graph. To remove the Average of SellingPrice, right-click on the gray box in the Pivot Chart that says Average of SellingPrice, then select Remove Field.
7. Repeat step 6 to remove the Average of ProcessTime.
8. Let's also remove the 2023 data because there is not an entire year of data for 2023. In the Pivot Chart, click the down arrow next to Years, and unselect 2023 and > 1/28/2023.
9. Let's also change Sum of SellingPrice to Sales Revenue by replacing "Sum of SellingPrice" with "Sales Revenue" in the pivot table heading. Sales revenue is the sum of all of the selling prices.
10. Let's also change Sum of GrossProfit to Gross Profit by replacing "Sum of GrossProfit" with "Gross Profit" in the pivot table heading.

The resulting pivot chart will be as shown in Lab Exhibit 7.2.9.

Lab Exhibit 7.2.9



You can use the drop-down arrow next to Size to select specific table sizes.

Assessment

1. Take a screenshot of the pivot table for total sales revenue, average selling price, average process time, and total gross profit by table size, year, and quarter, copy it into a

Word document named "Lab 7.2 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the pivot chart for total sales revenue and total gross profit by table size, year, and quarter; copy it into the same Word document; and label the screenshot Submission 2.

2. Answer the questions in Connect and upload your Lab 7.2 Excel Submission.docx via Connect, if assigned.

Alternate Lab 7.2 Excel

Descriptive Analytics: Using a Pivot Table to Analyze Historical Performance by Product Size and Year

Apply the same steps in Lab 7.2 to the Alt Lab 7.2 Data.xlsx data set.

Required

1. Create a pivot table for total sales revenue, average selling price, average process time, and total gross profit by table size, year, and quarter.
2. Create a pivot chart for total sales revenue and total gross profit by table size, year, and quarter.

Total Gross Profit: Gross profit for a table is sales price minus the cost of making the table. Total gross profit sums all of the gross profits for tables sold.

Specify the Question

What are total sales, average selling price, average process time, and total gross profit by table size, year, and quarter?

Assessment

1. Take a screenshot of the pivot table for total sales revenue, average selling price, average process time, and total gross profit by table size, year, and quarter; copy it into a Word document named "Alt Lab 7.2 Excel Submission.docx"; and label the screenshot Submission 1. Take a screenshot of the pivot chart for total sales revenue and total gross profit by table size, year, and quarter; copy it into the same Word document; and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Alt Lab 7.2 Excel Submission.docx via Connect, if assigned.

Lab 7.3 Tableau

Descriptive Analytics: Using a Histogram to Evaluate Process Time

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Descriptive Analytics, Tableau, Histogram

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the following questions "What happened?" and "What is happening?" Lab 7.1 used Excel's **Insert > Charts** to generate a histogram for marketing data. In this lab, we use Tableau to generate a histogram.

Several years ago, Rob built a table for his son to use in building LEGO creations. The table allowed his son to stand while building, provided a frame to hold LEGO base plates, and got the LEGOs off the floor. Several of Rob's neighbors saw the table and asked him to build tables for their children. The tables grew in popularity and Rob eventually quit his job in 2018 and founded LeTable Inc. He now works full-time building custom LEGO tables.

Customers select from four different sizes of tables:

1. Small, which holds 4 base plates
2. Medium, which holds 8 base plates
3. Large, which holds 16 base plates
4. Deluxe, which has multiple levels as well as conduits for electric lights

Customers can also select from five different colors:

1. Blue
2. Red
3. Yellow
4. Natural (polyurethane only)
5. Stained

Rob would like to see an analysis of the time to make and ship different table sizes (2018–2022).

Required

1. Create a histogram for process time.
2. Create an annual time series graph for average process time by table size.

Process Time: The time it takes to fill an order. It is measured as the number of days between receiving the order and shipping the table to the customer.

Specify the Question

How many days does it take to make and ship different table sizes?

Obtain the Data

To compare the histogram generated using the Histogram option of the Excel Analysis ToolPak, Excel's Insert Charts, and a histogram generated by Tableau (in this lab), this lab uses the same data as in Labs 7.1 and 7.2. The file, Lab 7.3 Data.xlsx, contains the replicated data. The data dictionary for this file is on the following page. Note that the order dates range from 1/5/2018 to 12/31/2022 for a total of 19,995 orders. In addition, some orders were not completed and shipped until 2023.

Data Dictionary

OrderID:	Sequential number identifying the sales order
CustomerName:	First and last name of the customer[35]
ZipCode:	5-digit zip code of the customer
OrderDate:	Date that the order was placed in MM/DD/YYYY format
ShipDate:	Date that the completed order was shipped in MM/DD/YYYY format
Color:	Color of the table (blue, red, yellow, natural, stained)
Size:	Table size (small, medium, large, deluxe)
SellingPrice:	Price the table was sold for
Cost:	Cost to make the table

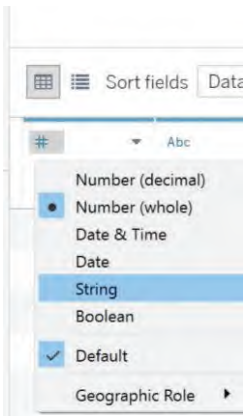
Obtain Task 1: Load the Data into Tableau.

1. Open Tableau.
2. Under Connect, open Microsoft Excel.
3. In the Open window, browse to the Lab 7.3 Data.xlsx file. Click Open. Notice that Tableau automatically separated the field names into separate words when a capital letter was used in the Excel column headings (thus, OrderID became Order ID).

Obtain Task 2: Prepare the Data for Analysis.

1. Tableau defined Order ID as a number (see the symbol # in the upper-left corner above Order ID). Because Order ID is an identifier/dimension, not a number used in computations, change Order ID to a string (). To do this, click on the # above Order ID and select String, as shown in Lab Exhibit 7.3T.1. Order ID should now be identified as an "Abc" type of data.

Lab Exhibit 7.3T.1



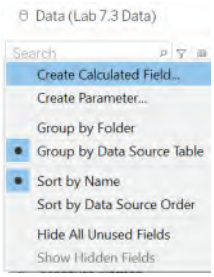
Analyze the Data

Analysis Task 1: Calculate Process Time.

1. Click on Sheet 1.
2. First, create a calculated field for Process Time, which is the number of days it took to make and ship the table after the order was placed (Ship Date – Order Date). Click the down arrow that appears to the right of the Search box in the Data pane and select Create Calculated Field, …, as shown in Lab Exhibit 7.3T.2.

[35]For Labs 7.1, 7.2, 7.3, 7.4, and 7.5, all names were generated using https://www.databasetestdata.com/.

Lab Exhibit 7.3T.2



3. In the Calculation Editor that appears, type "Process Time" as the title. Then in the Calculator Editor, type the following formula: "[Ship Date] – [Order Date]". The Calculation Editor should look like Lab Exhibit 7.3T.3. Note the message at the bottom: The calculation is valid. Click OK. A new calculated field called Process Time should be added to Measures in the Data pane (below the gray line). Calculated fields are identified with a "=#".

Lab Exhibit 7.3T.3

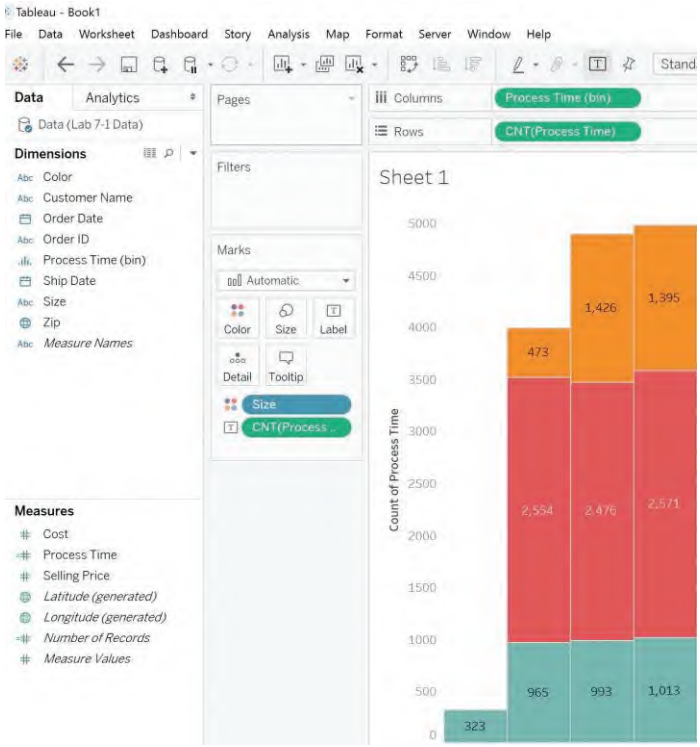


Analysis Task 2: Create a Histogram.


DATA VISUALIZATION

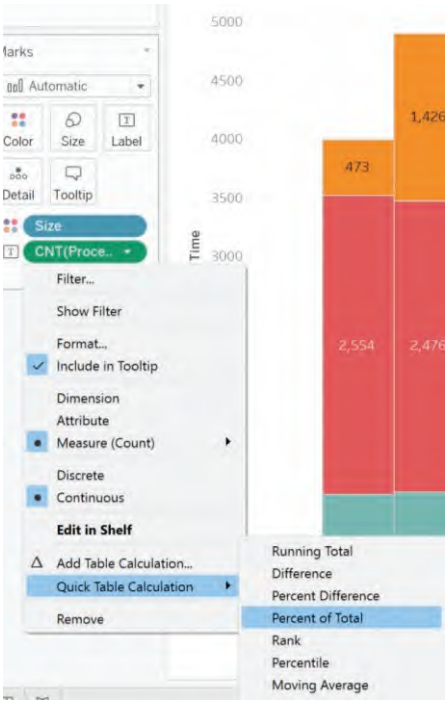
1. Drag Process Time from the Data pane to the Columns shelf. Click Show Me on the toolbar if it is not already visible. Then select the histogram chart ( ). A histogram should appear with the Columns shelf changed to Process Time (bin) and the Rows shelf changed to CNT(Process Time). Unlike Excel, Tableau can create its own bins based on the data. (*Note:* User-defined bin size is also an option, but it is not explored in this lab.) Process Time (bin) represents the number of days to ship, and Count of Process Time is the number of tables.
2. Drag Size from the Data pane to Color in the Marks card. The histogram should clearly show the various processing times for different table sizes. A legend should also appear on the right. If the Show Me window is on top of the legend, click on Show Me to make it disappear (that is, move behind).
3. To show the number of different table sizes making up each bar, first hold down the Ctrl key and drag CNT(Process Time) from the Rows shelf to Label in the Marks card. Holding down the Ctrl key creates a copy of the field (leaving the field in its original location). Lab Exhibit 7.3T.4 shows the first four columns of the histogram containing the count data.

Lab Exhibit 7.3T.4



4. To convert the count data to percentages, right-click on the CNT(Process Time) field in the Marks card. Select **Quick Table Calculation > Percent of Total,** as shown in Lab Exhibit 7.3T.5.

Lab Exhibit 7.3T.5



The numbers should have changed to percentages. So, 0.91 percent of all tables shipped 30–32 days after they were ordered (and they were all Deluxe tables).

5. Double-click on Sheet 1 at the bottom of the view and type "Process Time" to name your sheet. Click Enter.
6. Change the X- and Y-axis to more descriptive names. Double-click on the Y-axis Count of Process Time. In the Edit Axis dialog box, change the title to "Number of Orders". Click the X at the top right of the dialog box. Repeat for the X-axis, changing Process Time (bin) to "Days". To format the Y-axis as "X,XXX", right-click on the Y-axis and select Format. In the Axis tab on the left in the Scale section, set Numbers to Number (Standard).

Report the Results: Visualize Process Time over Time

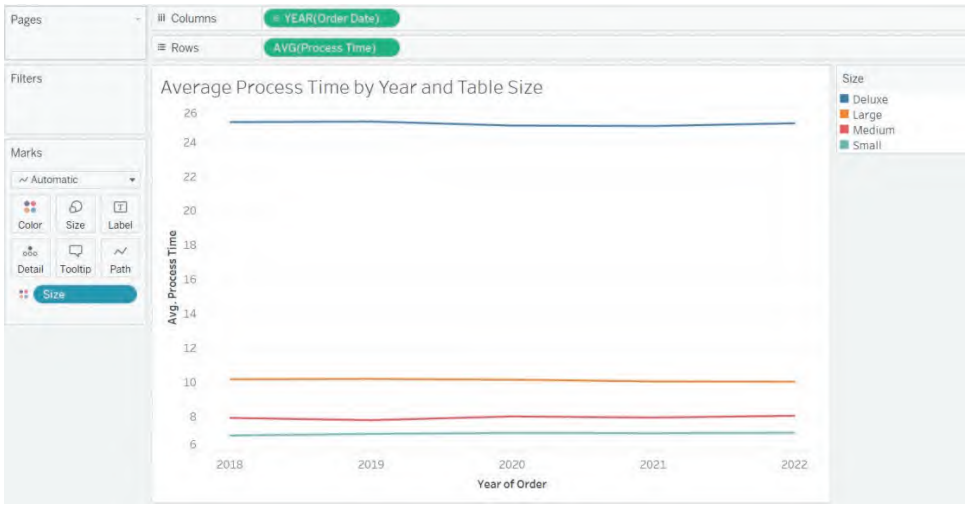


DATA VISUALIZATION

You have just created a histogram showing a breakdown of processing time for different sizes of tables. After looking at the visualization, we realize that another view of the data might be helpful to Rob. Conducting additional analysis often happens as you gain an understanding of the data. Specifically, let's create a new data visualization in Tableau presenting average processing time for each table size for each year in the data (2018–2022). To create this additional visualization, use the following steps:

1. Click on Sheet 2.
2. Drag Order Date to the Columns shelf. Click the down arrow in the pill and select the second Year choice.
3. Drag Process Time (not Process Time (bin)) to the Rows shelf.
4. Change the Process Time pill from SUM to AVG. Click the arrow in the pill, and select **Measure(Sum) > Average**.
5. Drag Size to Color in the Marks card.
6. Name the visualization "Average Process Time by Year and Table Size".
7. Right-click on the Y-axis, select Edit Axis, and uncheck the Include zero box in the General tab of the Edit Axis [Avg. Process Time] dialog box. Click x to close the dialog box. Rename the X-axis "Year of Order".

Lab Exhibit 7.3T.6 shows the new visualization: a line graph revealing that average process time for each table size has remained constant over time. Rob can use these data visualizations to determine if certain table sizes need reduced processing time. He can also use the processing time in conjunction with customer demand to determine which tables should be produced in what order.



Lab Exhibit 7.3T.6

Assessment

1. Take a screenshot of the completed histogram, paste it into a Word document named "Lab 7.3 Tableau Submission.docx", and label the screenshot Submission 1; Take a screenshot of the completed line graph, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Lab 7.3 Tableau Submission.docx file via Connect, if assigned.

Lab 7.3 Power BI

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Descriptive Analytics: Using a Histogram to Evaluate Process Time

Keywords

Descriptive Analytics, Power BI, Histogram

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" Lab 7.1 used Excel's **Insert > Charts** to generate a histogram for marketing data. In this lab, we use Power BI to generate a histogram.

Several years ago, Rob built a table for his son to use in building LEGO creations. The table allowed his son to stand while building, provided a frame to hold LEGO base plates, and got the LEGOs off the floor. Several of Rob's neighbors saw the table and asked him to build tables for their children. The tables grew in popularity and Rob eventually quit his job in 2018 and founded LeTable Inc. He now works full-time building custom LEGO tables.

Customers select from four different sizes of tables:

1. Small, which holds 4 base plates
2. Medium, which holds 8 base plates
3. Large, which holds 16 base plates
4. Deluxe, which has multiple levels as well as conduits for electric lights

Customers can also select from five different colors:

1. Blue
2. Red
3. Yellow
4. Natural (polyurethane only)
5. Stained

Rob would like to see an analysis of the time to make and ship different table sizes (2018–2022).

Process Time: The time it takes to fill an order. It is measured as the number of days between receiving the order until shipping the table.

Required

1. Create a histogram for process time.
2. Create an annual time series graph for average process time by table size.

Specify the Question

How many days does it take to make and ship different table sizes?

Obtain the Data

To compare the histogram generated using the Histogram option of the Excel Analysis ToolPak, Excel's Insert Charts, and a histogram generated by Power BI (in this lab), this lab uses the same data as in Labs 7.1 and 7.2. The file, Lab 7.3 Data.xlsx, contains the replicated data. The data dictionary for this file is below. Note that the order dates range from 1/5/2018 to 12/31/2022 for a total of 19,995 orders. In addition, some orders were not completed and shipped until 2023.

Data Dictionary

OrderID:	Sequential number identifying the sales order
CustomerName:	First and last name of the customer[36]
ZipCode:	5-digit zip code of the customer
OrderDate:	Date that the order was placed in MM/DD/YYYY format
ShipDate:	Date that the completed order was shipped in MM/DD/YYYY format
Color:	Color of the table (blue, red, yellow, natural, stained)
Size:	Table size (small, medium, large, deluxe)
SellingPrice:	Price the table was sold for
Cost:	Cost to make the table

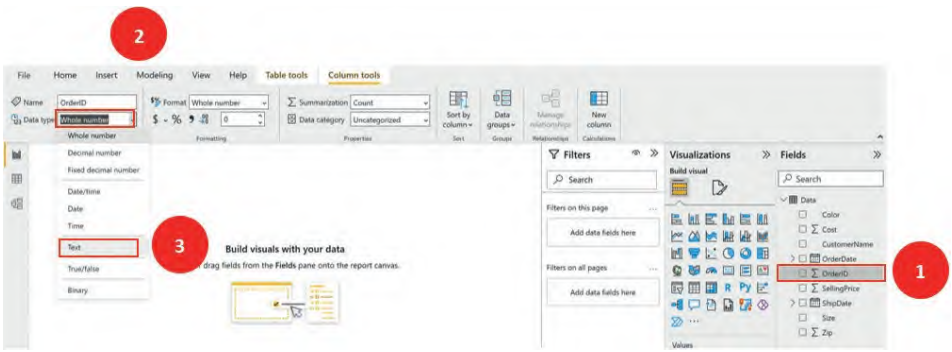
Obtain Task 1: Load the Data into Power BI.

1. Open Power BI.
2. Click Get Data.
3. In the Get Data dialog box, select Excel Workbook. Click Connect.
4. In the Open window, browse to the Lab 7.3 Data.xlsx file.
5. Click Open.
6. In the Navigator dialog box, place a check mark next to Data and click Load.

Obtain Task 2: Prepare the Data for Analysis.

1. Power BI defined Order ID as a number (see the sigma symbol next to the field in the Fields pane). Because Order ID is an identifier/dimension, not a number used in computations, change Order ID to a string. To do this, select the Order ID field from the Fields pane and adjust the Data type from Whole Number to Text, as shown in Lab Exhibit 7.3P.1.

Lab Exhibit 7.3P.1



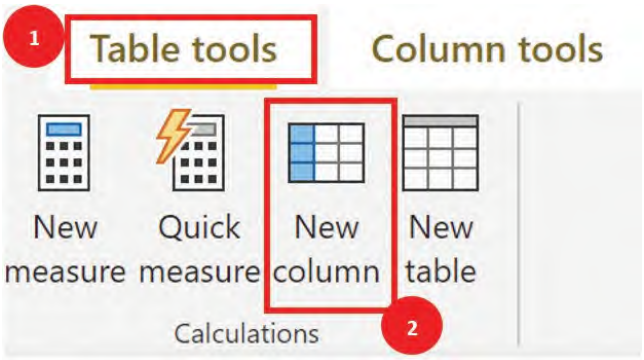
[36]For Labs 7.1, 7.2, 7.3, 7.4, and 7.5, all names were generated using https://www.databasetestdata.com/.

Analyze the Data

Analysis Task 1: Calculate Process Time.

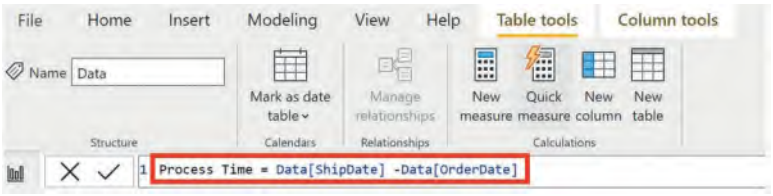
1. To determine the process time, which is the number of days it took to make and ship the table after the order was placed (Ship Date – Order Date), click the Table tools tab on the ribbon and select New Column, as shown in Exhibit 7.3P.2.

Lab Exhibit 7.3P.2



2. In the field that appears, type "Process Time = Data[ShipDate]–Data[OrderDate}". If a dropdown window appears after [OrderDate], click the blinking | cursor to make it close. Click the check mark. See Lab Exhibit 7.3P.3.

Lab Exhibit 7.3P.3

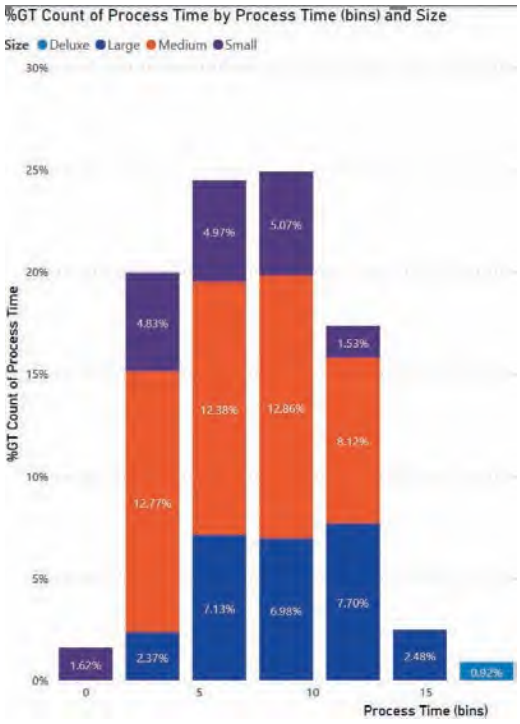


3. The new field Process Time will default to a Date/time Data type. Adjust the data type to a Decimal number in the Column tools tab.

DATA VISUALIZATION

Analysis Task 2: Create a Histogram.

1. First, create bins for the histogram. Right-click the Process Time field in your Fields pane and select New group.
2. Change the Bin size to 2.95, then click OK.
3. Next, click the blank area on your page (where it says "Build visuals with your data"), and then click the Stacked column chart in the Visualizations pane (with the Build visual tab selected). Resize the chart to fill the page.
4. In the Visualizations pane, make sure you are in the Build visual tab. The  should be highlighted.
5. Drag Process Time (bins) from the Fields pane to the box under X-axis where it says "Add data fields here" in the Visualizations pane.
6. Drag Process Time from the Fields pane to the box under Y-axis where it says "Add data fields here" in the Visualizations pane, and adjust the aggregation from the default of Sum to Count using the v dropdown. In addition, using the v dropdown, change Show value as to: Percent of grand total.
7. Drag Size from the Fields pane to the box under Legend where it says "Add data fields here".
8. To show the number of different table sizes making up each bar, click on the Format your visual icon  in the Visualizations pane. Toggle/slide the Data labels option to On. Lab Exhibit 7.3P.4 shows the first seven columns of the histogram.
9. Rename the X-axis, Y-axis, and title. Click on the Format (your) visual icon and then click on the Visual tab. Find the X-axis. Click the > dropdown. Expand Title and change the Title Text box from Auto to "Days". Click the v to make the X-axis menu



options disappear. Repeat the process for the Y-axis, changing the Y-axis name from Auto to "Number of Orders". To change the chart title, click on the General tab (still in the Format visual option). Expand the Title and change the default Text from %GT Count of Process Time by Process Time (bins) and Size to "Process Time".

10. Add count to the Tooltip so that when you hover over a column it shows the number of tables. Click on the Build visual icon of the Visualizations pane. Drag Process Time from the Fields pane to the box under Tooltips where it says "Add data fields here" in the Visualizations pane, and adjust the aggregation from the default of Sum to Count using the v dropdown. Now when you hover over a column, a count indicating the number of tables should appear.

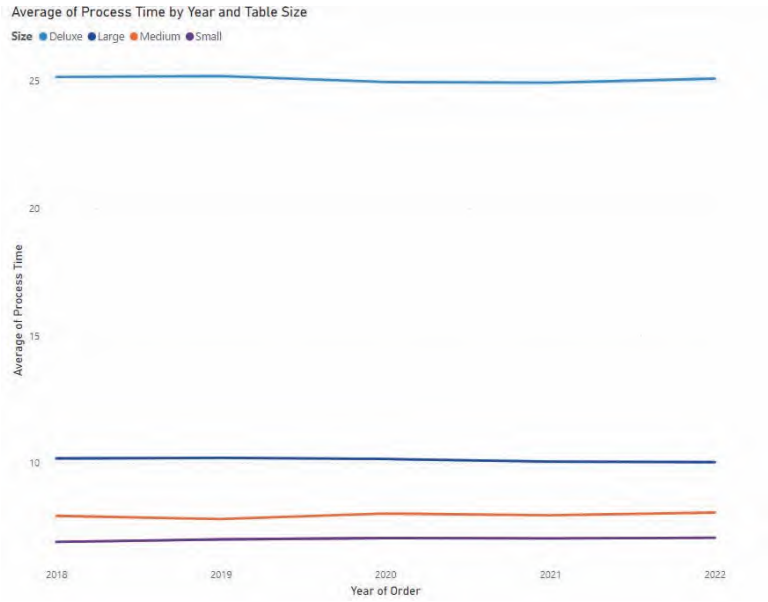
Report the Results: Visualize Process Time over Time

You have just created a histogram showing a breakdown of process time for different sizes of tables. Conducting additional analysis often happens as you gain an understanding of the data. After looking at the visualization, we realize that another view of the data might be helpful to Rob. So, let's create a new data visualization in Power BI presenting average processing time for each table size for each year in the data (2018–2022). To create this additional visualization, use the following steps:

1. Click on + at the bottom of the canvas to create Page 2.
2. Click the blank area on your page, and then click the Line chart in the Visualizations pane in the Build visual option. Resize the chart to fill the page.
3. Verify that the Build visual tab of the Visualizations pane is selected.
4. Drag Process Time (not Process Time (bins)) from the Fields pane to the to the box under Y-axis where it says "Add data fields here" in the Visualizations pane.
5. Change Process Time from Sum to Average using the v dropdown.
6. Drag Size from the Fields pane to the box under Legend where it says "Add data fields here".
7. Drag OrderDate from the Fields pane to the box under X-axis where it says "Add data fields here" in the Visualizations pane.
8. Rename the title and X-axis following the directions in step 9 of Analysis Task 3. The title should be "Average Process Time by Year and Table Size". The X-axis title should be "Year of Order".

Lab Exhibit 7.3P.5 shows the new visualization: a line graph revealing that average process time for each table size has remained constant over time. Rob can use these data visualizations to determine if certain table sizes need reduced processing time. He can also use the processing time in conjunction with customer demand to determine which tables should be produced and in what order.

Lab Exhibit 7.3P.5



Assessment

1. Take a screenshot of the completed histogram, paste it into a Word document named "Lab 7.3 Power BI Submission.docx", and label the screenshot Submission 1. Take a screenshot of the completed line graph, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Lab 7.3 Power BI Submission.docx file via Connect, if assigned.

Alternate Lab 7.3

Descriptive Analytics: Using a Histogram to Evaluate Process Time

Using either Tableau or Power BI, apply the same steps in Lab 7.3 to the Alt Lab 7.3 Data.xlsx data set.

Required

1. Create a histogram for process time. If you are using Power BI, select 2.65 for the bin size.
2. Create an annual time series graph for average process time by table size.

Assessment

1. Take a screenshot of the completed histogram, paste it into a Word document named "Alt Lab 7.3 Submission.docx", and label the screenshot Submission 1. Take a screenshot of the completed line graph, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Alt Lab 7.3 submission.docx file via Connect, if assigned.

Lab 7.4 Excel

Diagnostic Analytics: Analyzing the Steps in the Sales Process with a Sales Funnel Chart

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Diagnostic Analytics, Excel, Sales Funnel Chart

Decision-Making Context

Diagnostic analytics helps to answer the questions "Why did it happen?" and "What are the causes of past results?" One diagnostic technique is a sales funnel chart, which shows the progressive reduction of data from one stage in a process to the next stage in the process. Marketers use sales funnel charts to visualize (1) the steps in the sales process and (2) how the number of potential customers and associated (potential) sales revenue decrease in each stage.

Several years ago, Rob built a table for his son to use in building LEGO creations. The table allowed his son to stand while building, provided a frame to hold LEGO base plates, and got the LEGOs off the floor. Several of Rob's neighbors saw the table and asked him to build tables for their children. The tables grew in popularity and Rob eventually quit his job in 2018 and founded LeTable Inc. He now works full-time building custom LEGO tables.

During the period from 2018 to 2022, LeTable Inc. made tables in a variety of sizes (small, medium, large, and deluxe) and colors (blue, red, yellow, natural, stained). Due to the increasing popularity of the product, Rob launched a new product at the beginning of 2023 in hopes of capturing an even bigger market. Customers can now order a completely custom table (in addition to the original product offerings). Rob also has created a monthly newsletter and started organizing LEGO competitions/events across the country. All product sales are made through a series of five steps on LeTable's website. In other words, LeTable Inc. has only one marketing channel.

Rob's website gathers five types of information, as follows:

1. Visitors: The total number of potential customers that visit LeTable's website.
2. Marketing qualified leads (MQLs): Potential customers who signed up for LeTable's new monthly newsletter, which highlights custom-built tables and LEGO builds by customers.
3. Sales qualified leads (SQLs): Potential customers signing up for information about local LEGO competitions/events. Customers must have signed up for the newsletter first to receive information about the competitions/events.
4. Quotes: Potential customers who asked for a price quote for a table.
5. Closed sales: Customers who purchased a table.

Rob has MQL signup website data for the first quarter of 2023. Data also show the website visitors who then continued to other stages of buying a table by providing an email address, signing up for local LEGO competitions (SQLSignUP), asking for a quote (QuoteDate), and placing an order (OrderDate). Rob wants to analyze the website data with a funnel chart.

Required

Create a sales funnel chart in Excel for LeTable's marketing process using website data.

Specify the Question

How many customers are in each of the sales process steps for LeTable Inc.? What percentage of customers who ask for a quote on a table make a purchase?

Obtain the Data

This lab uses the data in Lab 7.4 Data.xlsx. The file contains two worksheets. The worksheet labeled "Sales Process Steps" contains the five types of information listed above along with the number of website visitors in the first quarter of 2023 (a total of 97,832 visitors), which Rob identified using **Google Analytics**. The remaining fields will be completed by summing the data in the Data worksheet. The Data worksheet contains the items defined in the data dictionary below.

Data Dictionary

Email:	(Potential) customer email address
MQLSignUpDate:	Date (potential) customers provide email to receive monthly newsletter
SQLSignUpDate:	Date (potential) customers sign up for local LEGO competitions/events
QuoteDate:	Date when quote was provided to (potential) customer based on customer specifications for table
OrderDate:	Date that the customer placed the order

Analyze the Data

Analysis Task 1: Set Up the Spreadsheet.

1. Click on the Data worksheet. Note that it contains a list of emails with the dates for signing up to receive the newsletter (MQLSignUpDate), signing up for LEGO competition/event information (SQLSignUpDate), sending a quote for a table (QuoteDate), and placing an order (OrderDate). Not all customers complete all steps, as indicated by blank spaces. We need to count the potential customers in each step and place that number in the appropriate Sales Process Steps worksheet cell.
2. Click on the Sales Process Steps worksheet. Notice that Rob has already filled in the number of website visitors during the quarter from Google Analytics (97,832). The Source to Calculate in columns G and H is designed to help you identify which information should be used to complete Column C of the table. We will use the Excel function COUNTA to count the non-empty cells in the B–E columns on the Data worksheet. Type the following into the designated cells in the Sales Process Steps worksheet:

   C3: =COUNTA(Data!$B$2:$B$26453)
   C4: =COUNTA(Data!$C$2:$C$26453)
   C5: =COUNTA(Data!$D$2:$D$26453)
   C6: =COUNTA(Data!$E$2:$E$26453)

Your spreadsheet should look like Lab Exhibit 7.4.1.

Lab Exhibit 7.4.1

	A	B	C	D	E	F	G	H
1	Step	Step Description	Quantity	Percent of Prior Sales Process Step	Percent of Website Visitors		Source to Calculate	Data worksheet column
2	1	Website Visitors	97,832				Google Analytics (given)	
3	2	Marketing Qualified Leads (MQLs)	26,452				MQLSignUpDate on Data worksheet	B
4	3	Sales Qualified Leads (SQLs)	9,112				SQLSignUpDate on Data worksheet	C
5	4	Quotes	5,227				QuoteDate on Data worksheet	D
6	5	Closed Sales	3,971				OrderDate on Data worksheet	E

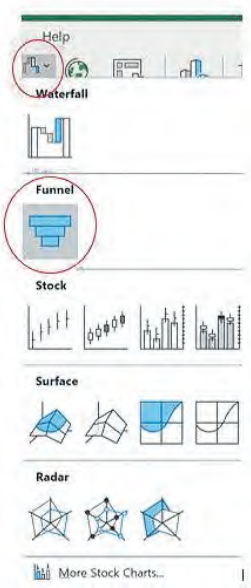
3. Next, we need to calculate the percentages in cells D3–D6. In cell D3, type "=C3/C2" and format the cell using the % icon on the Home tab. This step indicates the percentage of website visitors who become MQLs.
4. Copy the formula in cell D3 down to cell D6. These cells determine the percentage of MQLs who become SQLs, the percentage of SQLs who ask for a quote, and the percentage of quotes that become actual orders, respectively.
5. Finally, we need to calculate the percentages in cells E2–E6. In cell E2, type "=C2/$C$2" and format the cell using the % icon on the Home tab.
6. Copy the formula in cell E2 down to cell E6. This column shows the percent of website visitors who took part in each step of the process.

Analysis Task 2: Create a Funnel Chart.



1. In the Sales Process Steps worksheet, highlight cells B2:C6.
2. Select **Insert > Charts > Insert Waterfall, Funnel, Stock, Surface, or Radar Chart** in Charts. Click on Funnel, as shown in Lab Exhibit 7.4.2.

Lab Exhibit 7.4.2



3. The funnel chart should automatically appear. Resize the chart so that you can see all of the numbers.
4. Click on the Chart Title to change the title to "Website Funnel Chart 1st Quarter 2023".

Report the Results: Create a Visually Stunning Sales Funnel Chart



You have just created a sales funnel chart showing the number of (potential) customers at each step in the website sales process. However, the sales funnel chart automatically created by Excel looks pretty basic. You would like a more professional-looking funnel chart to give to Rob, who plans to make a presentation to potential investors. Excel does have options to create a more attractive chart, but it will take a little more work than the ready-to-go funnel chart. Follow these steps to create a presentation-ready funnel chart.

1. In the Sales Process Steps worksheet, copy cells B1 through C6.
2. Paste the copied data into a new worksheet tab using Paste Special Values.

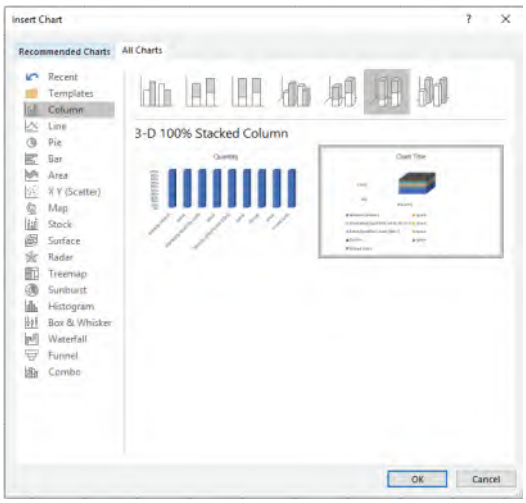
3. Insert rows between the data using "space" as the Step Description and "10000" as Quantity, as shown in Lab Exhibit 7.4.3. The spaces are used to create a gap between the steps and will be deleted in subsequent steps.

Lab Exhibit 7.4.3

Step Description	Quantity
Website Visitors	97,832
space	10000
Marketing Qualified Leads (MQLs)	26,452
space	10000
Sales Qualified Leads (SQLs)	9,112
space	10000
Quotes	5,227
space	10000
Closed Sales	3,971

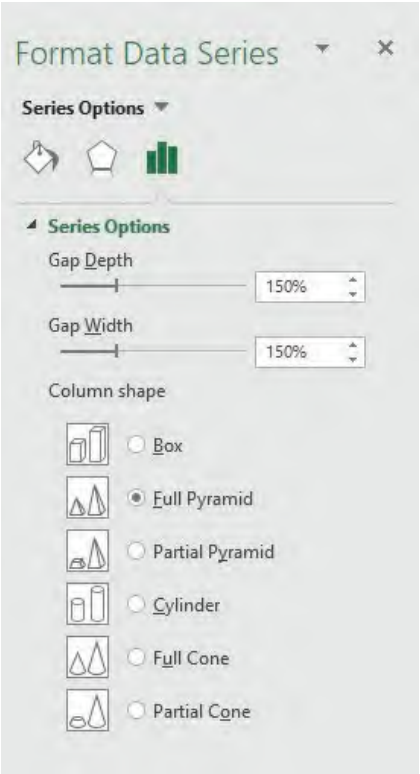
4. Highlight cells A1:B10 on the new worksheet tab.
5. From the menu, select **Insert > Charts > Insert Column or Bar Chart > More Column Charts.**
6. Select 3-D 100% Stacked Column (the sixth graph from the left).
7. Select the picture of the cube, as shown in Lab Exhibit 7.4.4. Click OK.

Lab Exhibit 7.4.4



8. Make both the chart box and the cube as large as possible.
9. Right-click on the Sidewall or vertical axis (the 0-100% tower on the left). A rectangular box around the 0-100% should appear. Select Format Axis.... Check Values in reverse order in the Format Axis dialog box. Doing so puts the number of website visitors on top.
10. Right-click on the cube. Select Format Data Series.... Then select Full Pyramid, as shown in Lab Exhibit 7.4.5. (*Note:* To change the shape, change Gap Depth and Gap Width.)

Lab Exhibit 7.4.5



11. To delete the lines around the pyramid, right-click on a line and select Delete.
12. To delete the percentages, right-click on the sidewall or vertical axis of percentages and select Delete.
13. To create the spaces, click on each of the space colors so blue circles appear, right-click, and set Fill to No Fill, as shown in Lab Exhibit 7.4.6. Four spaces should be set to No Fill. This creates an appearance of the different stages floating.

Lab Exhibit 7.4.6

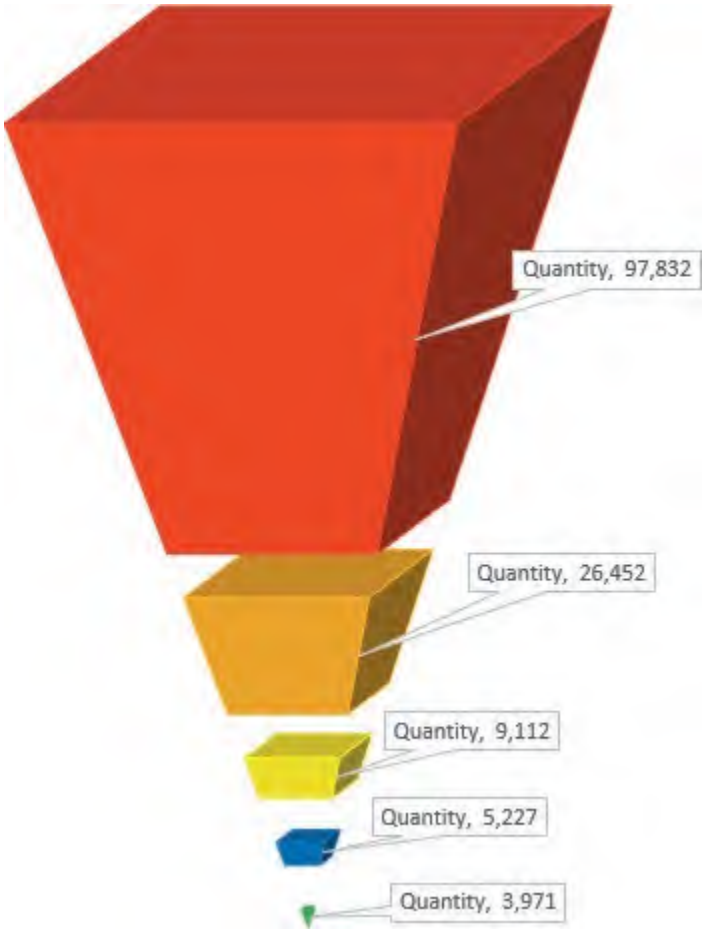


14. Let's format the funnel chart. Delete Quantity. To delete the four space entries in the legend, double-click on each space in the legend so four blue circles appear, right-click, and select Delete. Change Chart Title to "LeTable Inc. Website Funnel Chart 1st Quarter 2023". You can also change colors of each part of the funnel as well as add desired callouts in any color (**Chart Tools > Chart Layouts > Add Chart**

Element > Data Labels > Data Callouts). You can also change the size of the spaces between the layers by changing the size of the space rows, and you can add additional labels (for example, percentage of website visitors for each step, percentage of prior step).

Your funnel chart should look something like Lab Exhibit 7.4.7. Colors and callouts may vary according to your preference.

Lab Exhibit 7.4.7



Creating it takes some work, but the custom-made funnel chart will impress an audience. Rob can use the funnel charts to help decide where to focus his marketing efforts to help (potential) customers make it to the bottom of the funnel.

Assessment

1. Take a screenshot of the completed Sales Process Steps worksheet tab, copy it into a Word document named "Lab 7.4 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the Excel generated funnel chart, copy it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the other funnel chart, copy it into the same Word document, and label the screenshot Submission 3.
2. Answer the questions in Connect and upload your Lab 7.4 Excel Submission.docx file via Connect, if assigned.

Alternate Lab 7.4 Excel

Diagnostic Analytics: Analyzing the Steps in the Sales Process with a Sales Funnel Chart

Apply the same steps in Lab 7.4 to the Alt Lab 7.4 Data.xlsx data set. Rob has already added the number of website visitors from **Google Analytics** (109,463).

Required

Create a sales funnel chart in Excel for LeTable Inc.'s marketing process using website data.

Assessment

1. Take a screenshot of the completed Sales Process Steps worksheet tab, copy it into a Word document named "Alt Lab 7.4 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the Excel-generated funnel chart, copy it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the other funnel chart, copy it into the same Word document, and label the screenshot Submission 3.
2. Answer the questions in Connect and upload your Alt Lab 7.4 Excel submission.docx file via Connect, if assigned.

Lab 7.5 Tableau

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Diagnostic Analytics: Examining Pricing Strategy with Cluster Analysis

Keywords

Diagnostic Analytics, Tableau, Cluster Analysis

Decision-Making Context

Diagnostic analytics helps to answer the questions "Why did it happen?" and "What are the causes of past results?" One diagnostic technique is cluster analysis. Companies use cluster analysis to group like or similar customers into clusters. They can then use this information to create a plan for achieving more effective customer marketing for each cluster.

Several years ago, Rob built a table for his son to use in building LEGO creations. The table allowed his son to stand while building, provided a frame to hold LEGO base plates, and got the LEGOs off the floor. Several of Rob's neighbors saw the table and asked him build tables for their children. The tables grew in popularity and Rob eventually quit his job in 2018 and founded LeTable Inc. He now works full-time building custom LEGO tables in a variety of sizes and colors.

Seeing Rob's hard work and subsequent success, his sister, Maria, started her own business, called LeLights. LeLights sells custom lights for LEGO builds that integrate into LeTables. Maria has collected sales data for her first year of business. Unlike Rob, who uses a set selling price with a known cost per table (except for the fully custom tables in Lab 7.4), Maria's lights vary in selling price and cost based on the customer's specifications (type, size, and number). In addition, she has negotiated different discounts with different clients to increase her total sales revenue. While her first year was good (strong total sales and positive total gross profit), Maria wonders if she should continue her existing pricing strategy with occasional discounts.

Total Gross Profit: Gross profit for a table is sales price minus the cost of making the table. Total gross profit sums all of the gross profits for tables sold.

Required

Cluster LeLights' customer data by selling price and gross profit.

Specify the Question

Should LeLights change its pricing strategy?

Obtain the Data

The sales data are in the Lab 7.5 Data.xlsx file. The Data worksheet contains the items defined in the data dictionary below.

Data Dictionary[37]

OrderID:	Sequential number identifying the sales order
CustomerName:	First and last name of the customer
Discount:	Discount on total order
ZipCode:	5-digit zip code of the customer
State:	State customer lives in
Region:	Region customer lives in (Central, South, West)
GrossProfit:	Dollar amount made on the sales after subtracting product costs from sales
SellingPrice:	Total dollar amount paid by customer for the product, net of sales discount

[37]The data used in Lab 7.5 are based on the Superstore Data supplied with Tableau (slightly modified).

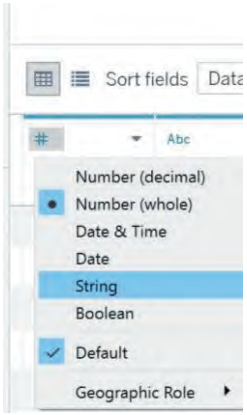
Obtain Task 1: Load the Data into Tableau.

1. Open Tableau.
2. Under Connect, select Microsoft Excel.
3. In the Open window, browse to the Lab 7.5 Data.xlsx file.
4. Click Open. Notice that Tableau automatically separated the field names into separate words when a capital letter was used in the Excel column headings (so OrderID became Order ID).

Obtain Task 2: Prepare the Data for Analysis.

1. Tableau defined Order ID as a number (see the symbol # in the upper-left corner above Order ID). Because Order ID is an identifier/dimension, not a number used in computations, change Order ID to a string (). To do this, click on the # above Order ID and select String, as shown in Lab Exhibit 7.5.1. Order ID should now be identified as an "Abc" type of data.

Lab Exhibit 7.5.1



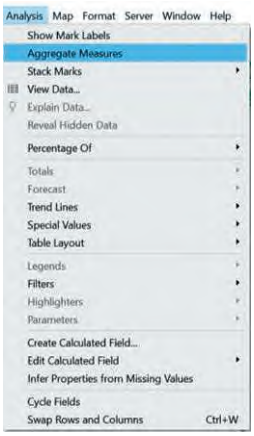
Analyze the Data

Analysis Task 1: Start the Visualization.


DATA VISUALIZATION

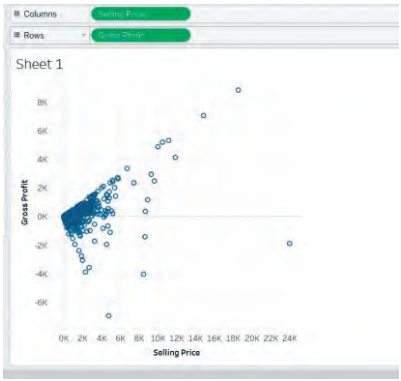
1. Click on Sheet 1.
2. Drag Selling Price from the Data pane to the Columns shelf.
3. Drag Gross Profit from the Data pane to the Rows shelf. Only one circle appears because the default in Tableau is to sum, or aggregate, the variables. (Note that you can tell total sales revenue and total gross profit for the year from this visualization.) We can use the Tableau menu to show the individual data points. In the Analysis menu option, unselect Aggregate Measures, as shown in Lab Exhibit 7.5.2.

Lab Exhibit 7.5.2



The visualization should look like Lab Exhibit 7.5.3. The visualization shows that some of the sales have a negative Gross Profit, indicating that the Selling Price (after sales discount) did not cover the cost to make the items.

Lab Exhibit 7.5.3



Analysis Task 2: Perform a Cluster Analysis.

You are now ready to perform a cluster analysis.

1. Click on the Analytics tab (next to the Data pane tab on the left).
2. Drag Cluster over the canvas. A Create Cluster dialog box will appear.
3. Drag Cluster over the Cluster box, as shown in Lab Exhibit 7.5.4.

Lab Exhibit 7.5.4

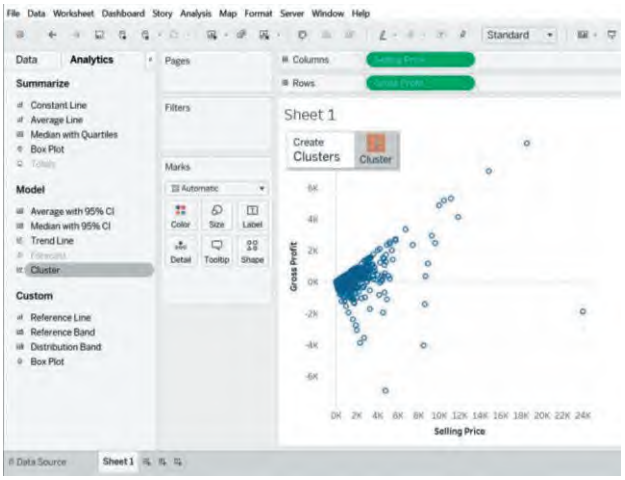


Tableau automatically clusters your data into groups of similar characteristics based on the variables on the canvas (here, Selling Price and Gross Profit). Notice that Tableau created three color-coded clusters. Observations in one cluster are more alike than observations in other clusters. Specifically, Tableau minimizes the distances between individual points in the cluster and the cluster center using an iterative procedure that is beyond the scope of this textbook.

Analysis Task 3: Explore the Cluster Analysis.

1. To see observations in just one cluster, click on the desired cluster in the Clusters legend (if Show Me is in the way, click the Show Me tab and it should go into the background). In addition, the Clusters dialog box appears. You can use this box to add variables (drag from the Data pane) or subtract variables (drag from the Clusters dialog box) from your cluster analysis. The Clusters dialog box also has a field for

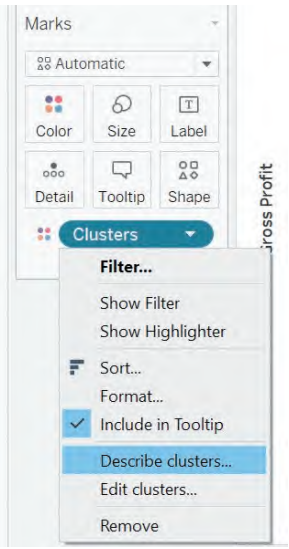
Number of Clusters, which defaults to Automatic, indicating that Tableau determined the number of clusters. You can also specify a specific number of desired clusters by typing in the number. Lab Exhibit 7.5.5 shows what the visualization should look like.



Lab Exhibit 7.5.5

2. To see the details of the clusters, click on the down arrow in the Clusters pill (in the Marks card). Select Describe clusters. . . , as shown in Lab Exhibit 7.5.6. (*Note:* You can use Edit clusters. . . to get the Clusters dialog box back.)

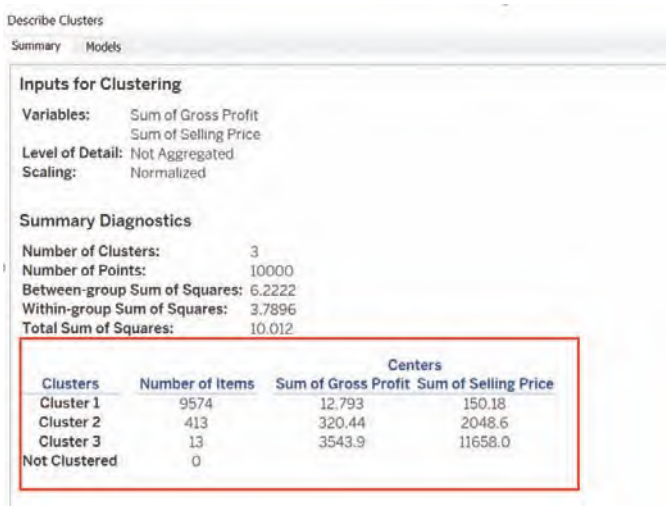
Lab Exhibit 7.5.6



3. The information outlined in red in Lab Exhibit 7.5.7 provides information about the clusters using Gross Profit and Selling Price. Centers provides the average value within each cluster. So, Cluster 1 has the lowest Gross Profit and Selling Price, Cluster 3 has the highest Gross Profit and Selling Price, and Cluster 2 falls in the middle.

Also, notice that most of the sales are in Cluster 1, indicating that most items are not making a high Gross Profit.

Lab Exhibit 7.5.7



Describe Clusters

Summary | Models

Inputs for Clustering

Variables:	Sum of Gross Profit Sum of Selling Price
Level of Detail:	Not Aggregated
Scaling:	Normalized

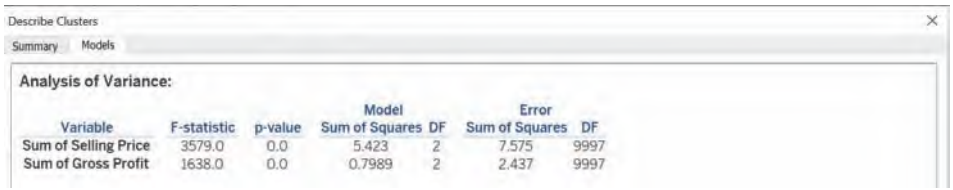
Summary Diagnostics

Number of Clusters:	3
Number of Points:	10000
Between-group Sum of Squares:	6.2222
Within-group Sum of Squares:	3.7896
Total Sum of Squares:	10.012

		Centers	
Clusters	Number of Items	Sum of Gross Profit	Sum of Selling Price
Cluster 1	9574	12.793	150.18
Cluster 2	413	320.44	2048.6
Cluster 3	13	3543.9	11658.0
Not Clustered	0		

4. To see the statistics, click on the Models tab. As Lab Exhibit 7.5.8 shows, the Models tab shows the two variables used for clustering and analyzes which variables are the best for creating clusters. Smaller *p*-values (typically less than 0.05) indicate that the variable values differ between the clusters. So, both Gross Profit and Selling Price differ between the clusters, indicating that the three clusters differ on those two characteristics (variables). (*Note:* Both the Summary and Models tabs change when you change the variables used to perform a cluster analysis.)

Lab Exhibit 7.5.8



Describe Clusters

Summary | Models

Analysis of Variance:

			Model		Error	
Variable	F-statistic	p-value	Sum of Squares	DF	Sum of Squares	DF
Sum of Selling Price	3579.0	0.0	5.423	2	7.575	9997
Sum of Gross Profit	1638.0	0.0	0.7989	2	2.437	9997

5. Close the Describe Clusters dialog box.
6. Double-click on the Sheet 1 tab and type "Cluster Analysis" to name the sheet.



DATA VISUALIZATION

Report the Results: Make the Cluster Visualization Easier to Read

You have just performed a cluster analysis to help analyze the profitability of LeLights' sales. However, as you review the data visualization, you realize that it will be helpful to identify which sales correspond to which data point. Let's identify each data point using the following steps:

1. Drag Order ID to Detail in the Marks card. Now, when you hover over a specific data point, Tableau identifies the Order ID, making it easy to determine which sales were the most profitable and which ones were not.
2. Drag Customer Name to Detail in the Marks card. Now, when you hover over a specific data point, Tableau identifies the not only the Order ID but also the Customer Name, as shown in Lab Exhibit 7.5.9.



Lab Exhibit 7.5.9

This detailed information should help Maria evaluate her current pricing strategy and whether it should change.

Assessment

1. Take a screenshot of the completed cluster visualization, copy it into a Word document named "Lab 7.5 Tableau Submission.docx", and label the screenshot Submission 1. Take screenshots of the cluster analysis tabs in the Describe Clusters dialog box, and copy them into the same Word document. Label the Summary tab Submission 2 and label the Models tab Submission 3.
2. Answer the questions in Connect and upload your Lab 7.5 Tableau Submission.docx via Connect, if assigned.

Alternate Lab 7.5 Tableau

Diagnostic Analytics: Examining Pricing Strategy with Cluster Analysis

Begin with your analysis from Lab 7.5, in which you performed a cluster analysis using Gross Profit and Selling Price. (You will continue using the data in Lab 7.5 Data.xlsx.) Add an additional variable, Discount, to the Clusters dialog box. If the Clusters dialog box is not showing, click on the down arrow in the clusters pill in the Marks card, and select Edit Clusters. . . .

Required

Cluster LeLights' customer data by Selling Price, Gross Profit, and Discount.

Assessment

1. Take a screenshot of the completed cluster visualization, copy it into a Word document named "Alt Lab 7.5 Tableau Submission.docx", and label the screenshot Submission 1. Take screenshots of the cluster analysis tabs in the Describe Clusters dialog box, and copy them into the same Word document. Label the Summary tab Submission 2 and label the Models tab Submission 3.
2. Answer the questions in Connect and upload your Alt Lab 7.5 Tableau Submission.docx via Connect, if assigned.

Lab 7.6 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Predictive Analytics: Predicting Sales Revenue from Advertising Expense

Keywords

Predictive Analytics, Excel, Regression, Advertising, Sales Revenue

Decision-Making Context

Predictive analytics attempts to answer the questions "Will it happen in the future?" and "What is the probability something will happen?" and "Can we forecast what will happen?" The goal is to predict future performance using historical data. So, unlike descriptive and diagnostic analytics, which describe things we know are true, predictive analytics describes things that possibly might happen. One predictive technique is regression, which predicts a specific outcome based on the relationship between inputs and outputs. Using the data, regression calculates an equation that has a generic form of y = intercept + mx, where y is the dependent (outcome) variable, x is the independent variable, and m is the slope of the line. For a refresher on regression, please see Chapter 3.

Companies can use regression to see if revenues are related to advertising expense. Companies spend money on advertising because the more (potential) customers know about or are reminded of a product and its qualities, the more product they supposedly will buy. However, the impact of advertising isn't always clear.

LeTable Inc., which we analyzed in Labs 7.1 through 7.5, creates custom tables for LEGOs. The tables hold base plates for big LEGO builds and can be customized to the customer's specifications. Over the last several years, the founder/owner, Rob, has started spending money on advertising to increase sales via the company's website (which is LeTable Inc.'s only marketing channel). Rob would like to know if the company's increased advertising expense is increasing sales revenue.

One way to determine the relationship is to run a regression using advertising expense as the input (independent variable) and sales revenue as the outcome (dependent variable).

Required

1. Using regression, determine the relationship between advertising expense and sales revenue.
2. Predict future sales revenue for a given level of advertising expense.

Specify the Question

What is the relationship between advertising expense and sales revenue? Given a specific level of advertising expense, what will sales revenue be?

Obtain the Data

This lab uses the data in Lab 7.6 Data.xlsx. The file contains advertising expense and sales revenue over the past 36 months for LeTable Inc. Here is the data dictionary for this file:

Data Dictionary

Month:	The analysis month number (1–36)
Date:	The beginning of the month date when sales revenues and advertising expense are recorded
Advertising Expense:	The monthly advertising expense for LeTable Inc.
Sales Revenue:	The monthly sales revenues for LeTable Inc.

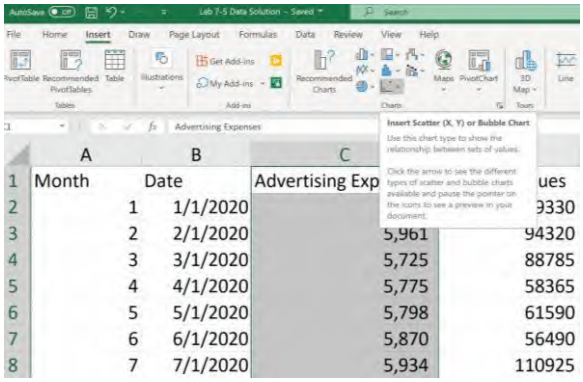
Analyze the Data



Analysis Task 1: Graph the Relationship.

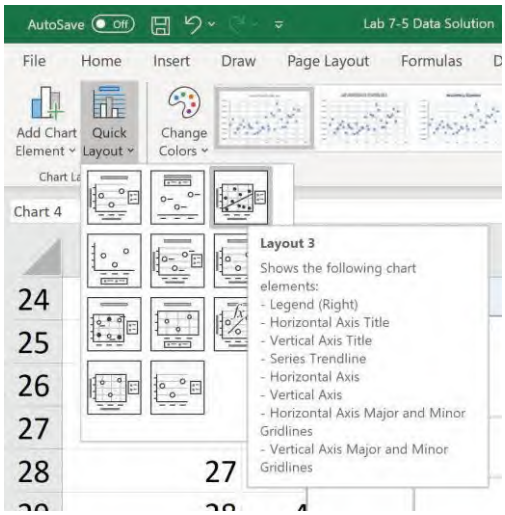
1. Let's start by plotting the data. The plot will allow us to see if there is a relationship between the variables. Highlight cells C1:D37.
2. Select **Insert > Charts > Scatter (X, Y) or Bubble Chart** from the menu bar, as shown in Lab Exhibit 7.6E.1.

Lab Exhibit 7.6E.1



3. Select the first choice (Scatter).
4. When the Chart Design Tab appears, select **Chart Design > Quick Layout > Layout 3,** as shown in Lab Exhibit 7.6E.2.

Lab Exhibit 7.6E.2



5. To label the X-axis as "Advertising Expense" and the Y-axis as "Sales Revenue", double-click on the desired Axis Title and type the appropriate name. The result is a graph plotting sales revenue and advertising expense, as shown in Lab Exhibit 7.6E.3. (*Note:* If you need to format the axes numbers to 0 decimals, click on the axis you want to format, right-click and select Format Axis. . . , select Number, set Decimal places to 0, and click outside the Format Axis dialog box.)

Lab Exhibit 7.6E.3



Why did we put advertising expense on the X-axis? Companies generally spend money on advertising to increase sales, so advertising expense might be a predictor of sales revenue. In other words, sales revenue may be a function of advertising expense, a relationship that we express formally as:

$$\text{Sales Revenue} = f(\text{Advertising Expense})$$

Given the graph in Lab Exhibit 7.6E.3, does there appear to be a relationship between advertising expense and revenue? If so, does the relationship appear to be positive? (In other words, as advertising expense increases, do sales increase?)

Analysis Task 2: Run the Regression.

Regression analysis uses statistics to determine (1) if variables have an impact on an outcome and (2) the size of the impact. Regression analysis helps answer the questions "Which factors matter the most in predicting an outcome?" and "Which factors can we ignore (because they don't predict the outcome)"?

Let's think through the relationship first. For example, it seems that the greater the advertising expense, the more likely sales revenue will increase (as that's the goal of advertising). Therefore, we expect a positive relationship between advertising expense and sales revenue, which means the regression line should move upwards to the right.

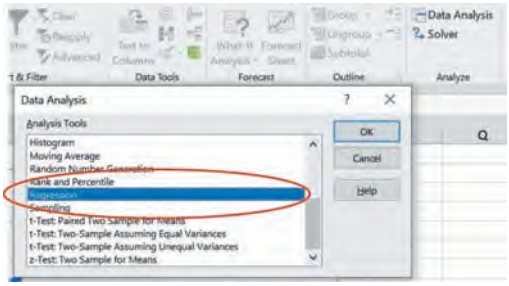
To see if there is a statistical relationship between advertising expense and sales revenue, we will run a regression with the following specifications:

- Dependent (outcome) variable: Sales revenue
- Independent variable: Advertising expense

To run the regression, we use the Excel Analysis ToolPak (see textbook Appendix E for instructions on how to install). Make sure the ToolPak is installed by selecting the Data tab. If Data Analysis is shown in the Analyze group, you are ready to go. Run the regression by following the steps below.

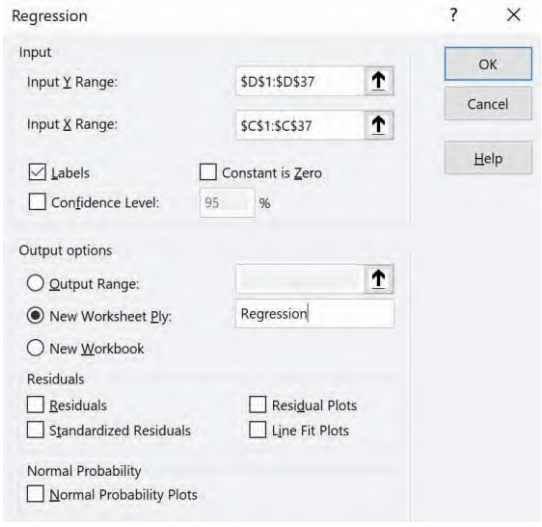
1. Select **Data > Analyze > Data Analysis.**
2. In the Data Analysis dialog box, select Regression, as shown in Lab Exhibit 7.6E.4. Click OK.
3. The Regression dialog box appears. Enter the dependent variable, or Input Y Range. The dependent (outcome) variable is Sales Revenue, so type "$D$1:$D$37" in the Input Y Range box.

Lab Exhibit 7.6E.4



4. Enter the independent variable, or Input X Range. The independent variable is Advertising Expense, so type "$C$1:$C$37" in the Input X Range box.
5. Click the Labels box, as the input ranges included the labels in $C$1 and $D$1.
6. Select New Worksheet Ply and type "Regression" in the box to name the new worksheet.
   The completed Regression dialog box should look like Lab Exhibit 7.6E.5.

Lab Exhibit 7.6E.5



7. Click OK to get the regressions results shown in Lab Exhibit 7.6E.6.

Lab Exhibit 7.6E.6

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.938303428							
5	R Square	0.880413323							
6	Adjusted R Square	0.876896068							
7	Standard Error	7482.208163							
8	Observations	36							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	14013360598	1.4E+10	250.3126	3.0175E-17			
13	Residual	34	1903436926	55983439					
14	Total	35	15916797524						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	-739681.3309	54672.0907	-13.5294	3.03E-15	-850788.387	-628574.2747	-850788.387	-628574.2747
18	Advertising Expense	146.6608032	9.269849791	15.82127	3.02E-17	127.8222019	165.4994045	127.8222019	165.4994045

Analysis Task 3: Analyze the Regression.

1. Let's look at the fit of the overall model. The regression with Advertising Expense fits the data better than an intercept-only model (that is, a model with no independent variables) because the number in the "Significance F" column is very small.

2. The Adjusted R Square equals 0.88. Adjusted R Square is another measure used to estimate how well the model predicted the dependent variable or output (Sales Revenue). The Adjusted R Square has a value between 0 and 1. An Adjusted R Square value of 0 indicates that the model has no ability to explain the dependent variable, while an Adjusted R Square value of 1 indicates that the model perfectly explains the dependent (outcome) variable. In this lab, an Adjusted R Square of 0.88 represents a high ability to explain the level of sales revenues given advertising expense.
3. Let's look at the significance of the individual independent variable. The t Stat for Advertising Expense (15.82) has a small P-Value (3.02E-17), which indicates that there is a significant positive relationship between Advertising Expense and Sales Revenue. In other words, Sales Revenue increases with higher Advertising Expense.
4. The regression coefficients are used to determine the regression equation (or fitted line plot). The regression equation for this data set is:

$$y = -739{,}681 + 146.66x$$

We can use this regression equation to predict Sales Revenue. For example, assuming that LeTable Inc. plans to spend $10,000 on advertising, predicted sales are $726,919:

$$\$726{,}919 = -\$739{,}681 + \$(146.66 * 10{,}000).$$

Report the Results

You ran a regression to determine the relationship between advertising and sales. Rob has not provided you with any estimated advertising expense, but he has asked you to predict future sales revenue for a given level of advertising expense of at least $10,000. We can use the regression equation to make those predictions. Let's estimate sales for five levels of advertising expense: $10,000, $12,500, $15,000, $17,500, and $20,000. In addition, we will set up the spreadsheet to make future predictions easy.

1. Add an additional tab to the spreadsheet file called Predictions.
2. Type "Advertising Expense" in cell A1.
3. Type "Projected Sales" in cell B1.
4. In cells A2–A6, type the advertising expense amounts: 10,000, 12,500, 15,000, 17,500, and 20,000.
5. In cell B2, type the following formula: "=−739,681 + 146.66*A2".
6. Copy the formula down the column to cell B6.
7. Format columns A and B as dollars with no decimals.
8. From the menu, select **Home > Styles > Format as a Table > *youpick*** to format the columns as a table of your own selection.
9. Make any additional formatting changes that you would like.

You have created a table showing predicted Sales Revenue based on a variety of Advertising Expense levels. The scatterplot, regression equation, and table of predicted Sales Revenue should help Rob decide how much to spend on advertising.

Assessment

1. Take a screenshot of the completed scatterplot, copy it into a Word document named "Lab 7.6 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the regression summary output, copy it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the prediction table and label the screenshot Submission 3.
2. Answer the questions in Connect and upload your Lab 7.6 Excel Submission.docx via Connect, if assigned.

Lab 7.6 Tableau

Predictive Analytics: Predicting Sales Revenue from Advertising Expense

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Predictive Analytics, Trend Line, Tableau, Regression, Advertising, Sales Revenue

Decision-Making Context

Predictive analytics attempts to answer the questions "Will it happen in the future?" and "What is the probability something will happen?" and "Can we forecast what will happen?" The goal is to predict future performance using historical data. So, unlike descriptive and diagnostic analytics, which describe things we know are true, predictive analytics describes things that possibly might happen. One predictive technique is regression, which predicts a specific outcome based on the relationship between inputs and outputs. Companies can use regression to see if revenues are related to advertising expense. Companies spend money on advertising because the more (potential) customers know about or are reminded of a product and its qualities, the more product they supposedly will buy. However, the impact of advertising isn't always clear.

LeTable Inc., which we analyzed in Labs 7.1 through 7.5, creates custom tables for LEGOs. The tables hold base plates for big LEGO builds and can be customized to the customer's specifications. Over the last several years, the founder/owner, Rob, has started spending money on advertising to increase sales via the company's website (which is LeTable Inc.'s only marketing channel). Rob would like to know if the company's increased advertising expense is increasing sales revenue.

One way to determine the relationship is to run a regression using advertising expense as the input (independent variable), and sales revenue as the outcome (dependent variable). To show the relationship between advertising expense and sales revenue, you can use a trend line, which you can create in Tableau. A trend line connects two or more data points, showing the movement or change in a measure. When creating a trend line, Tableau uses regression to calculate an equation that has a generic form of y = intercept + mx, where y is the dependent (outcome) variable, x is the independent variable, and m is the slope of the line. For a refresher on regression, please see Chapter 3. Lab 7.6 Excel asked you to run a regression in Excel. This lab has you add a trend line to a data visualization in Tableau.

Required

1. Using a trend line, determine the relationship between advertising expense and sales revenue.
2. Predict future sales revenue for a given level of advertising expense.
3. Write a formal report of your analysis.

Specify the Question

What is the relationship between advertising expense and sales revenue? Given a specific level of advertising expense, what will sales revenue be?

Obtain the Data

To help you compare Excel's regression with Tableau's trend line, this lab uses the same data as in Lab 7.6 Excel: Lab 7.6 Data.xlsx. The file contains advertising expense and

sales revenue over the past 36 months for LeTable Inc. Here is the data dictionary for this file:

Data Dictionary

Month:	The analysis month number (1–36)
Date:	The beginning of the month date when sales revenues and advertising expense are recorded
Advertising Expense:	The monthly advertising expense for LeTable Inc.
Sales Revenue:	The monthly sales revenues for LeTable Inc.

Obtain Task 1: Load the Data into Tableau.

1. Open Tableau.
2. In the Connect pane, click Microsoft Excel. Browse to the Lab 7.6 Data.xlsx file and click Open. The data should be ready to use.

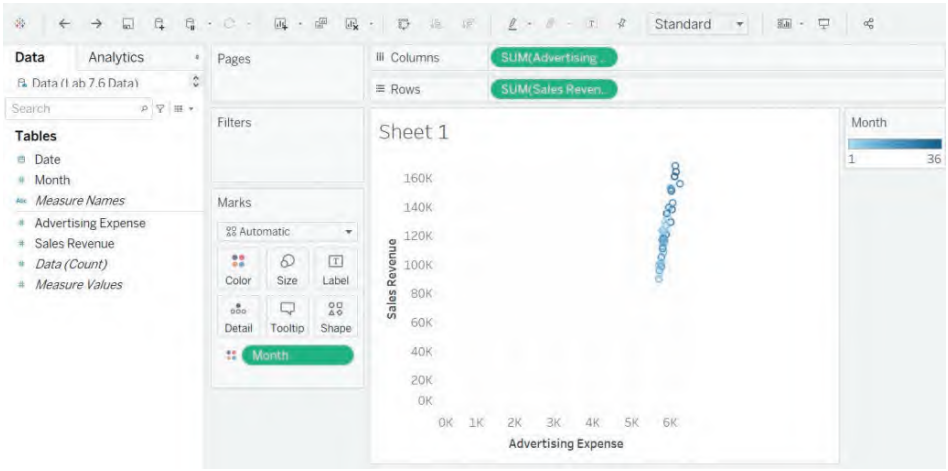
Analyze the Data



Analysis Task 1: Create a Scatterplot.

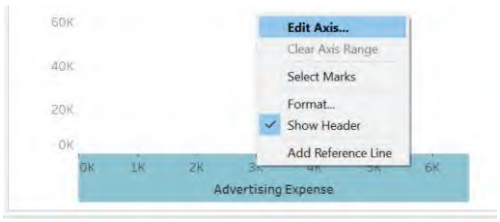
1. Click on Sheet 1.
2. Drag Advertising Expense to the Columns shelf.
3. Drag Sales Revenue to the Rows shelf.
4. Drag Month to Color in the Marks card. Your Tableau should look like Lab Exhibit 7.6T.1.

Lab Exhibit 7.6T.1



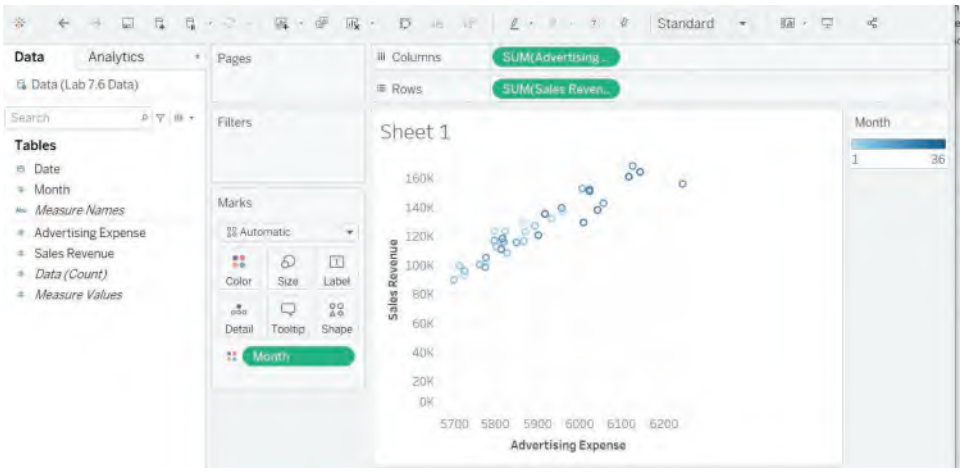
5. To make the visualization easier to interpret, let's have the X-axis not start at 0, because the data start above 5,000. Right-click on Advertising Expense (on the X-axis). Select Edit Axis. . . , as shown in Lab Exhibit 7.6T.2.

Lab Exhibit 7.6T.2



6. Uncheck the Include zero box. Click the X to close the Edit Axis dialog. Your Tableau should look like Lab Exhibit 7.6T.3.

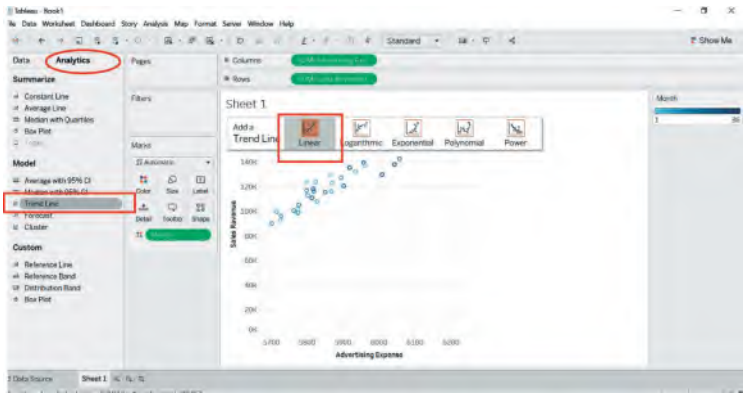
Lab Exhibit 7.6T.3



Analysis Task 2: Add a Trend Line.

1. Select the Analytics pane (the tab next to the Data pane tab).
2. Drag Trend Line onto the scatterplot and the Add a Trend Line dialog box will appear. Drop the Trend Line on the Linear model type, as shown in Exhibit 7.6T.4.

Lab Exhibit 7.6T.4

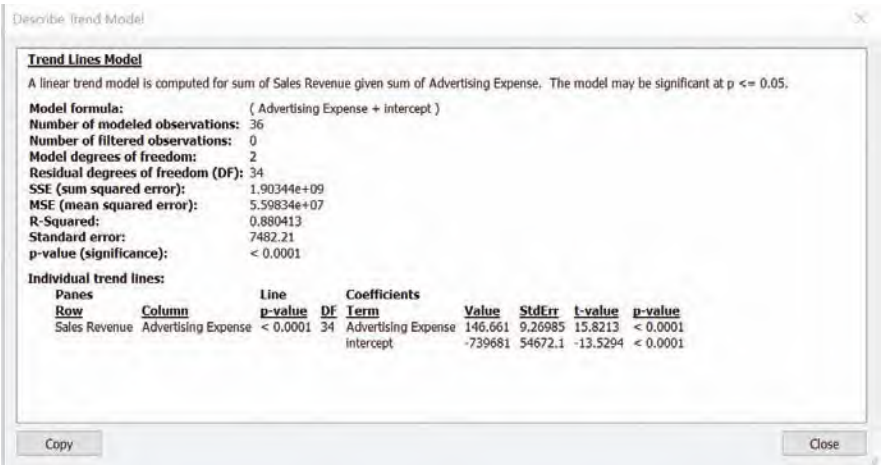


3. To name the worksheet Advertising – Sales Revenue, click on the Sheet 1 tab, type "Advertising – Sales Revenue", and click Enter.
4. To add commas to the X- and Y-axes, right-click on the X- or Y-axis and select Format... in the Axis tab, and then set Numbers equal to Number (Standard).

Analysis Task 3: Analyze the Trend Line.

1. Move the cursor over the Trend Line. Tableau shows the line equation as well as the R-Squared and P-Value. If you want more information, right-click on the Trend Line, and select Describe Trend Model. The Describe Trend Model dialog box opens and looks like Lab Exhibit 7.6T.5.

Lab Exhibit 7.6T.5


Describe Trend Model

Trend Lines Model

A linear trend model is computed for sum of Sales Revenue given sum of Advertising Expense. The model may be significant at p <= 0.05.

Model formula:	(Advertising Expense + intercept)
Number of modeled observations:	36
Number of filtered observations:	0
Model degrees of freedom:	2
Residual degrees of freedom (DF):	34
SSE (sum squared error):	1.90344e+09
MSE (mean squared error):	5.59834e+07
R-Squared:	0.880413
Standard error:	7482.21
p-value (significance):	< 0.0001

Individual trend lines:

Panes		Line		Coefficients				
Row	Column	p-value	DF	Term	Value	StdErr	t-value	p-value
Sales Revenue	Advertising Expense	< 0.0001	34	Advertising Expense	146.661	9.26985	15.8213	< 0.0001
				Intercept	-739681	54672.1	-13.5294	< 0.0001

Copy Close

2. Examine the output in the Describe Trend Model dialog, which is shown in Exhibit 7.6T.5. First, the number of observations used (36) is reported. The R-Squared measure (0.88) indicates how well the model predicted the dependent (outcome) variable (Sales Revenue). The R-Squared has a value between 0 and 1. An R-Squared value of 0 indicates that the model has no ability to explain the dependent variable, while an R-Squared value of 1 indicates that the model perfectly explains the dependent variable.[38] Thus, an R-Squared equal to 0.88 represents a high ability to explain the level of sales revenues given advertising. The p-value (significance) for the overall regression model is the same that was reported as Significance F in Excel. It means that the trend line with Advertising Expense fits the data better than an intercept-only model (that is, a model with no explanatory variables).
3. The coefficients that appear at the bottom right of the Describe Trend Model dialog box provide information about Advertising Expense and the intercept. The t-value for Advertising Expense (15.8213) has a very small *p*-value (<0.0001), indicating a significant positive relationship between Advertising Expense and Sales Revenue. That is, Sales Revenue increases with additional Advertising Expense.
4. The regression coefficients are used to determine the regression equation (or fitted line plot). Here, the regression equation is:

$$y = -739,681 + 146.66x$$

We can use the regression line to predict Sales Revenue for a given Advertising Expense. For example, assume that LeTable Inc. plans to spend $10,000 on advertising. What is Sales Revenue expected to be? The answer is $726,919 = −$739,681 + $(146.66 * 10,000).

Report the Results: Create a Formal Report

You have just added a trend line to a Tableau visualization of Advertising Expense and Sales Revenue. Rob can use the trend line equation to decide how much he should spend on Advertising Expense in the upcoming months. You can now prepare a formal report for Rob that includes your analysis and an interpretation of the results. The scatterplot, regression equation, and report should help Rob decide how much to spend on advertising.

[38]Excel also includes the Adjusted R Square, which penalizes for insignificant variables in the equation and rewards for significant variables. R-Squared always increases. Adjusted R Square solves this problem.

Assessment

1. Save your formal report in a Word document named "Lab 7.6 Tableau Submission.docx".
2. Answer the questions in Connect and upload your Lab 7.6 Tableau Submission.docx via Connect, if assigned.

Alternate Lab 7.6

Predictive Analytics: Predicting Sales Revenue from Advertising Expense

Analyze the Alt Lab 7.6 Data.xlsx data set using Excel or Tableau. *Note*: If you are using Tableau, that you will also need to uncheck the Include zero box in the Edit Axis for Sales Revenue.

Required

1. Using regression or a trend line, determine the relationship between advertising expense and sales revenue.
2. Predict future sales revenue for a given level of advertising expense.
3. Create a formal report that includes your analysis and an interpretation of the results.

Assessment

1. Save your Excel or Tableau analysis in a Word document named "Alt Lab 7.6 Submission.docx".
2. Answer the questions in Connect and upload your Alt Lab 7.6 Submission.docx via Connect, if assigned.

Lab 7.7 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Prescriptive Analytics: Calculating Internet CPM Rate Using Goal Seek

Keywords

Prescriptive Analytics, Goal Seek, Excel, Cost-per-mille (CPM)

Decision-Making Context

Providing actionable intelligence is one of the goals of business analytics. Prescriptive analytics answers the questions "What should we do, based on what we expect will happen?" and "How do we optimize our performance based on potential constraints?" Prescriptive analytics recommends a course of action. This lab uses Excel's Goal Seek menu option to determine how much should be spent on internet advertising.

Digital marketing includes placing advertisements on search engines or carefully selected websites that are frequently visited by a company's target customers. To place an advertisement on a search engine or website, companies pay a fee. These fees are sources of revenue for search engines and websites.

As discussed in the chapter, fees can be charged based on the number of clicks or the number of views. Cost-per-click (CPC) fees are based on the number of clicks. In this fee structure, a company is charged only when a user clicks on the ad. For example, if the CPC is $0.25 per click and 1,000 users click on the ad, then the company will owe $250 to the sponsoring website. Cost-per-mille (CPM) fees are based on the number of impressions or views, counted in the thousands (*mille* means 1,000 in Latin). For example, a CPM of $150 means that the company would pay $150 for 1,000 views of the webpage containing the ad. CPM can also be subdivided into a per view cost. So, continuing with this CPM example, the cost of one view would be $150/1,000 equals $0.15 per view. The CPC fee structure typically converts to higher direct sales because only interested potential customers click, while the CPM fee structure ensures that a large number of potential customers see the ad, which is good for increasing brand or product awareness.

LeTable Inc., which we analyzed in earlier labs in this chapter, creates custom tables for LEGOs. The tables hold base plates for big LEGO builds and can be customized to the customer's specifications. The founder/owner, Rob, wants to purchase ads on **Rebrickable.com** and is about to negotiate the CPM rate. Rob's total marketing budget is $10,000. He wants to split the budget between print ads for LEGO competitions and internet ads, with 20 percent of the marketing budget going to print ads and 80 percent going to internet ads that are expected to have 8,000,000 views. He wants to know the CPM as well as the cost per view of the internet ads.

Required

1. Calculate the CPM rate.
2. Write a formal report of the results of the analysis.

Specify the Question

What is an acceptable CPM rate for internet advertising?

Obtain the Data

This lab uses the data in Lab 7.7 Data.xlsx. You will complete the yellow cells. Here is the data dictionary for this file.

Data Dictionary

Marketing Budget:	Total marketing budget for LeTable Inc.
Print Ads Percentage of Total Marketing Budget:	Percent of Marketing Budget spent on print ads
Print Ads Budget:	Total dollars spent on print ads (= Marketing Budget * Print Ads Percentage of Total Marketing Budget)
Internet Ads Percentage of Total Marketing Budget	Percent of Marketing Budget spent on internet ads
Internet Ads Budget:	Total dollars spent on internet ads (=Marketing Budget * Internet Ads Percentage of Total Marketing Budget;)
Cost per 1,000 views:	CPM rate charged by a website or search engine for 1,000 views of an ad
Cost per 1 view:	Rate charged by a website or search engine for 1 view of an ad
Total Views:	The total number of views of an ad on a website

Analyze the Data

Analysis Task 1: Calculate Variables.

1. The yellow cells need to be filled and formatted as described below.
2. In cell B1, type "10000". Format with a comma using **Home > Number > ,** from the menu.
3. In cell B3, type "0.2" because Rob wants to spend 20% of his marketing budget on print ads.
4. In cell B4, type "=B1*B3" to calculate the budget available for print ads. Format with a comma using **Home > Number > ,** from the menu.
5. In cell B6, type "0.8" because Rob wants to spend 80 percent of his marketing budget on internet ads.
6. In cell B7, type "=B1*B6" to calculate the budget available for internet ads. Format with a comma using **Home > Number > ,.**
7. In cell B8, type "=2.00". This is just a placeholder for now, as described below. This CPM amount is what we want to calculate.
8. In cell B9, type "=B8/1000" to calculate the cost per view, which is necessary to calculate the total number of views. Use the menu to format this cell to four decimal places using **Home > Number > Increase Decimal.** Four decimals is sometimes necessary to see the cost per 1 view.
9. In cell B10, type "=B7/B9" to calculate the total number of internet ad views. Format with a comma using **Home > Number > ,.**

Cell B8, CPM, is the value that Goal Seek will calculate. We've just put 2.00 as a beginning point. Excel will adjust the value in its calculations. Your spreadsheet should look like Lab Exhibit 7.7.1.

	A	B
1	Total Marketing Budget	10,000.00
2		
3	(1) Print Ads Percentage of Total Marketing Budget	0.2
4	Print Ads Budget	2,000.00
5		
6	(2) Internet Ads Percentage of Total Marketing Budget	0.8
7	Internet Ads Budget	8,000.00
8	Cost per 1,000 views (CPM)	2.00
9	Cost per 1 view	0.0020
10	Total Views	4,000,000.00

Lab Exhibit 7.7.1

Analysis Task 2: Perform a Goal Seek.

Goal Seek works backwards to determine the item of interest, the CPM rate in cell B8.

1. From the menu, select **Data > Forecast > What-if Analysis > Goal Seek.**
2. Type the following items into the boxes.

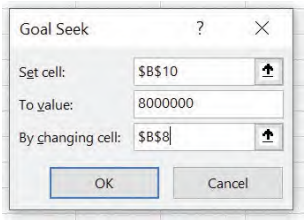
   Set cell: "$B$10"

   To value box: "8000000" (which is the number of views that Rob wants)

   By changing cell: "$B$8"

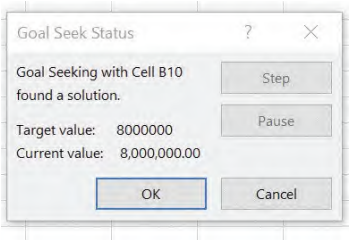
Lab Exhibit 7.7.2 presents the Goal Seek dialog box. It tells Excel to set Total views specified in cell $B$10 to Rob's desired viewership of 8,000,000 by changing the value of CPM, which is in cell $B$8. Basically, Excel's Goal Seek function works backwards to determine what CPM (or cost-per-mille) in cell B8 needs to be set at to achieve 8,000,000 views when the budget is $8,000.

Lab Exhibit 7.7.2



3. Click OK. Excel shows the iteration process as it runs. Once completed, the Goal Seek Status dialog box is shown as in Lab Exhibit 7.7.3.

Lab Exhibit 7.7.3



Cells B8, B9, and B10 should have been updated.

Report the Results: Write a Formal Report

You have just calculated the CPM rate that Rob should be willing to pay Rebrickable.com. Write a formal report in Word outlining your analysis, findings, and recommendations. The report should include a screenshot of the Excel worksheet results.

Assessment

1. Save your formal report as "Lab 7.7 Excel Submission.docx".
2. Answer the questions in Connect and upload your Lab 7.7 Excel Submission.docx via Connect, if assigned.

Alternate Lab 7.7 Excel

Prescriptive Analytics: Calculating Internet CPM Rate Using Goal Seek

Using the same data in Lab 7.7 Data.xlsx, use Excel to determine what the CPM rate should be if Rob wants to spend 70 percent of a $20,000 total marketing budget on an internet advertising campaign with at least 12,000,000 views.

Required

1. Calculate the CPM rate.
2. Write a formal report for the results of the analysis.

Assessment

1. Save your formal report as "Alt Lab 7.7 Excel Submission.docx".
2. Answer the questions in Connect and upload your Alt Lab 7.7 Excel Submission.docx via Connect, if assigned.

Lab 7.8 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Prescriptive Analytics: Calculating Product Price Using Goal Seek

Keywords

Prescriptive Analytics, Goal Seek, Excel, Product Pricing

Providing actionable intelligence is one of the goals of business analytics. Prescriptive analytics answers the questions: "What should we do, based on what we expect will happen?" and "How do we optimize our performance based on potential constraints?" Prescriptive analytics recommends a course of action. This lab uses Excel's Goal Seek menu option to determine the product price necessary to achieve a desired net income.

As described in the prior Chapter 7 labs, several years ago, Rob built a table for his son to use in building LEGO creations. The table allowed his son to stand while building, provided a frame to hold LEGO base plates, and got the LEGOs off the floor. Several of Rob's neighbors saw the table and asked him to build tables for their children. The tables grew in popularity and Rob eventually quit his job in 2018 and founded LeTable Inc. He now works full-time building custom LEGO tables in a variety of sizes and colors.

Lab 7.5 described how Rob's sister, Maria, created a successful business called LeLights, which sells custom lights for LEGO builds that integrate into LeTables. Now, seeing her dad's and aunt's success, Rob's teenage daughter, Isabella, wants to start a small business called Birthday Blocks that sells LEGO birthday creations. As an alternative to birthday cards, Isabella's birthday creations will be small, unique, personalized LEGO builds based on the recipient's interests and favorite colors. The builds can include pictures as well as robotics to incorporate movement. She wants to keep the business small for now because she is still in school. She would like to make a net income of $12,000 a year.

Isabella has estimated the demand (that is, the number of units) she has time to make and can sell each month. She estimated the unit cost to make each creation, assuming that she uses approximately the same cost of materials to make each build. She plans to keep the cost of the building materials low by buying used bricks and robotic parts at garage sales and flea markets. She will also spend $100 a month to rent a weekend booth at a weekly arts-and-craft fair where she will take and deliver orders. However, she is not sure what price to charge that customers would be willing to pay while also making her desired net income.

Required

1. Calculate the product price.
2. Create a column (vertical bar) chart for net income and cumulative net income by month.

Specify the Question

What price should Birthday Blocks charge for its product to make $12,000 net income?

Obtain the Data

The estimated units sold and unit variables costs are provided in Lab 7.8 Data.xlsx. You will complete the yellow cells. Here is the data dictionary for this file.

Data Dictionary

Price:	Price for each unique, personalized LEGO build
Units Sold:	Estimated demand for a month
Unit Variable Cost:	Cost to make one unique, personalized LEGO build
Total Sales Revenue:	The amount of money collected from customers during a month (Price * Units Sold)
Total Variable Cost:	Total cost for building all builds during a month (Units Sold * Unit Variable Cost)
Total Fixed Cost:	Monthly booth rental cost ($100)
Net Income:	The amount made in a month (= Sales Revenue – Total Variable Cost – Total Fixed Cost)

Analyze the Data

Analysis Task 1: Set Up the Spreadsheet.

1. The yellow cells need to be filled and formatted as described below. Note that Price is initially set to $10.00 but will be changed once we perform a Goal Seek. Units Sold is set at 60 units a month with a $4.00 variable cost. Also, note that rows 2, 3, and 4 have been set up so that the entry in column B controls the entries in columns C through M reflecting the assumption that price, units sold, and unit variable cost will be constant throughout the year.
2. In cell B6, type "=B2*B3" to determine total amount of money collected from customers during a month.
3. In cell B7, type "=B3*B4" to calculate the total variable costs of making the units during a month.
4. In cell B8, type "100" to record the fixed monthly rent of $100.
5. In cell B9, type "=B6-B7-B8" to calculate the net income for the month, which is equal to Total Sales Revenue minus Total Variable Cost minus Total Fixed Cost.
6. Copy cells B6, B7, B8, and B9 through column M.
7. Create five totals for the year.

   In cell N3, type "=SUM(B3:M3)" to determine total units sold during the year.
   In cell N6, type "=SUM(B6:M6)" to determine total sales revenue for the year.
   In cell N7, type "=SUM(B7:M7)" to determine total variable cost for the year.
   In cell N8, type "=SUM(B8:M8)" to determine total fixed cost for the year.
   In cell N9, type "=SUM(B9:M9)" to determine total net income for the year.

8. Using the menu, format rows 2, 4, 6, 7, 8, and 9 as dollars using **Home> Number > $.**

   Lab Exhibit 7.8.1 shows what your spreadsheet should look like.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		January	Febraury	March	April	May	June	July	August	September	October	November	December	Total
2	Price	$ 10.00	$ 10.00	$ 10.00	$ 10.00	$ 10.00	$ 10.00	$ 10.00	$ 10.00	$ 10.00	$ 10.00	$ 10.00	$ 10.00	
3	Units Sold	60	60	60	60	60	60	60	60	60	60	60	60	720
4	Unit Variable Cost	4	4	4	4	4	4	4	4	4	4	4	4	
5														
6	Total Sales Revenue	$ 600.00	$ 600.00	$ 600.00	$ 600.00	$ 600.00	$ 600.00	$ 600.00	$ 600.00	$ 600.00	$ 600.00	$ 600.00	$ 600.00	$ 7,200.00
7	Total Variable Cost	$ 240.00	$ 240.00	$ 240.00	$ 240.00	$ 240.00	$ 240.00	$ 240.00	$ 240.00	$ 240.00	$ 240.00	$ 240.00	$ 240.00	$ 2,880.00
8	Total Fixed Cost	$ 100.00	$ 100.00	$ 100.00	$ 100.00	$ 100.00	$ 100.00	$ 100.00	$ 100.00	$ 100.00	$ 100.00	$ 100.00	$ 100.00	$ 1,200.00
9	Net Income	$ 260.00	$ 260.00	$ 260.00	$ 260.00	$ 260.00	$ 260.00	$ 260.00	$ 260.00	$ 260.00	$ 260.00	$ 260.00	$ 260.00	$ 3,120.00

Lab Exhibit 7.8.1

Analysis Task 2: Perform a Goal Seek.

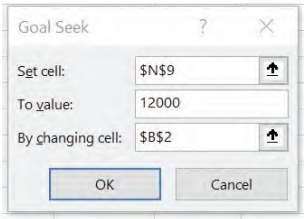
Goal Seek works backwards to determine the item of interest—specifically, the price in cell B2.

1. From the menu, select **Data > Forecast > What-if Analysis > Goal Seek.**
2. Type the following items into the boxes.

   Set cell: "$N$9" (Annual Net Income)
   To value box: "12000" (which is the Annual Net Income that Isabella wants)
   By changing cell: "$B$2" (item Price)

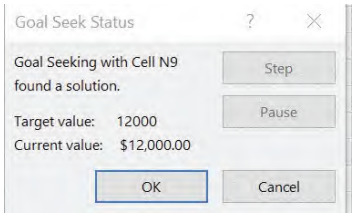
Lab Exhibit 7.8.2 presents the Goal Seek dialog box. It tells Excel to set the total net income for the year specified in cell $N$9 to Isabella's desired net income of $12,000 by figuring out the correct price to charge, which is in cell $B$2. Basically, Excel's Goal Seek function works backwards to determine what price in cell B2 is needed to earn $12,000 during the year (given the other estimates).

Lab Exhibit 7.8.2



3. Click OK. Excel shows the iteration process as it runs. Once completed, the Goal Seek Status dialog box is shown as in Lab Exhibit 7.8.3.

Lab Exhibit 7.8.3



The entire spreadsheet should have been updated using the calculated price.



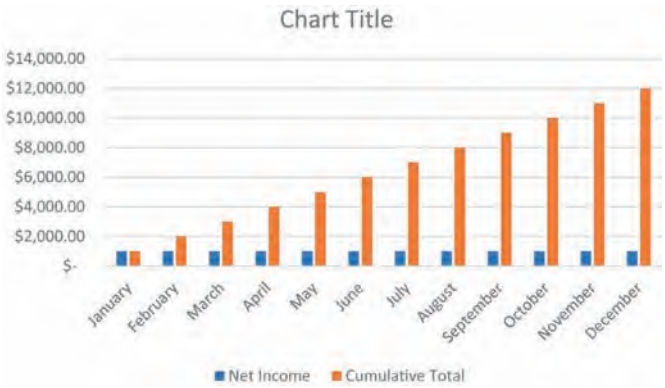
DATA VISUALIZATION

Report the Results: Visualize the Results

You have just calculated the price that Isabella should set for her birthday LEGO creations to earn $12,000 for the year. Let's create a visualization showing the monthly net income and cumulative net income for the year.

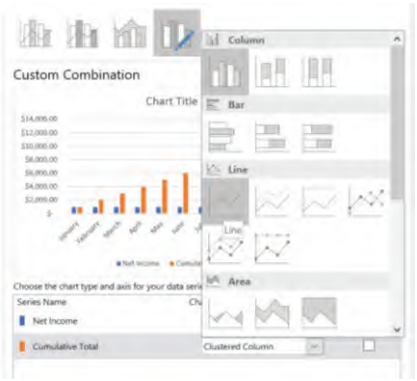
1. Create a new worksheet/tab in Excel titled Visualization.
2. In cells B1 through M1, type the months of the year beginning with January.
3. Copy the values in cells A9 through M9 in the Data tab to the new worksheet using Paste Special Values beginning in cell A2.
4. In cell A3, type "Cumulative Total".
5. In cell B3, type "=B2".
6. In cell C3, type "=C2 + B3".
7. Copy the formula in cell C3 to cells D3 through M3.
8. Place the cursor somewhere in the data.
9. From the menu, select **Insert > Charts > Recommended Charts.** From the All Charts tab, select Column, Clustered Column. (A quick shortcut uses Ctrl-F1 to bring up the recommended chart.) The chart looks like two clustered columns, as shown in Lab Exhibit 7.8.4.

Lab Exhibit 7.8.4



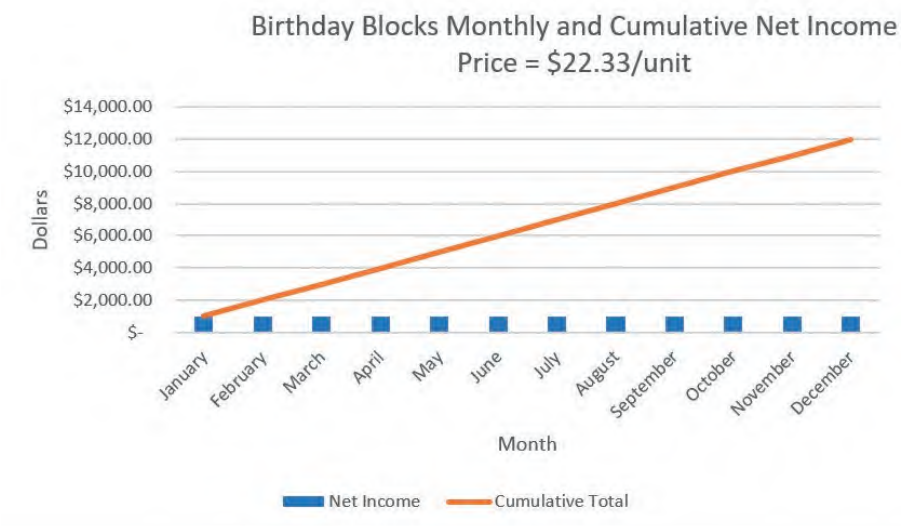
10. Click on an orange Cumulative Total bar so that all of the orange bars are highlighted with blue dots. Right-click, and then select Change Series Chart Type. Change the Cumulative Total Series to a Line Chart Type, as shown in Lab Exhibit 7.8.5.

Lab Exhibit 7.8.5



11. Change the Chart Title to "Birthday Blocks Monthly and Cumulative Net Income Price = $22.33/unit".
12. Add horizontal and vertical axis titles. In the Chart Tools Design tab, select **Add Chart Element > Axis Titles> Primary Horizontal.** To change the Axis Title, click on the Axis Title and type "Month". In the Chart Tools Design tab, select **Add Chart Element > Axis Titles> Primary Vertical.** Change the Axis Title to "Dollars". Your visualization should look like Lab Exhibit 7.8.6.

Lab Exhibit 7.8.6



Although this data visualization is simple, it is a powerful way to show two data series at the same time. The bars shown in this data visualization should help Isabella clearly understand how much money Birthday Blocks will earn monthly and over the year if the price is set at $22.33 per build.

Assessment

1. Take a screenshot of the monthly and cumulative income projections, copy it into a Word document named "Lab 7.8 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the data visualization, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Lab 7.8 Excel Submission.docx via Connect, if assigned.

Alternate Lab 7.8 Excel

Prescriptive Analytics: Calculating Product Price Using Goal Seek

Using data in the Alt Lab 7.8 Data.xlsx file, determine the product selling Price required to make $15,000. *Note:* Isabella has updated the monthly Units Sold to reflect the extra time she will have when school is not in session.

Required

1. Calculate the product price.
2. Create a column (vertical bar) chart for net income and cumulative net income by month.

Assessment

1. Take a screenshot of the monthly and cumulative income projections, copy it into a Word document named "Alt Lab 7.8 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the data visualization, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Alt Lab 7.8 Excel Submission.docx via Connect, if assigned.

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Chapter 8

Accounting Analytics

A Look Back

- Chapter 7 investigated how marketers can use business analytics by applying the SOAR model.
- Marketing helps companies meet customer needs profitably by selecting an appropriate marketing mix, which includes the four Ps of product, price, place, and promotion.

A Look at This Chapter

- Accounting is the systematic process of measuring, recording, and communicating the financial information that decision makers need to make decisions. Business analytics is increasingly used to answer accounting questions.
- Chapter 8 investigates how to use business analytics to address accounting questions by identifying accounting data sources, applying analytics techniques, and reporting results to decision-makers.

A Look Ahead

- Finance entails the management of money (or financial capital) by investing, borrowing, lending, budgeting, saving, and/or forecasting.
- Chapter 9 explains how to use business analytics to answer finance questions by identifying financial data sources, applying analytics techniques, and reporting results to decision-makers.

Digital innovations such as artificial intelligence, collaboration technologies, and advanced analytics are rapidly disrupting the finance 2020 workforce. Traditional roles are evolving, and newer roles are increasing in importance.

Traditional finance roles	Main impacts
Financial planning and analysis 	The real work starts when you deliver the reports or analysis: • Anticipating alternative scenarios, tracking their emergence, and executing on contingency plans • Not just answering the "what happened" and "why did it happen" questions but also answering the "what should we do" questions
Financial controller 	• Focusing on preventative and real-time control rather than relying on detective controls • Managing outcomes, not processes
Accounts payable clerk 	• Focusing on exceptions as routine work is automated • Greater collaboration with other functions

Analyst Skills Required by Today's Accounting Professionals

Accenture is a multinational professional services company that provides consulting services related to technology, business strategy, and operations. The company suggests that today's accounting and finance professionals need additional skills to handle modern technology and the mounds of data available to address accounting and finance questions.

Accenture emphasizes that the modern accounting and finance workforce must go beyond answering the "What happened?" and "Why did it happen?" questions with exploratory analytics. Workers must also answer the "Will it happen in the future?" and "What should we do?" questions using confirmatory analytics. They should also use prescriptive analytics to anticipate alternative scenarios before they occur. The exhibit above provides some examples of how digital innovations are affecting the modern workforce.

Source: "Meet the 2020 Workforce," Accenture, https://www.accenture.com/t20170511t060010z__w__/us-en/_acnmedia/pdf-50/accenture-finance-2020-workforce.pdf. Accessed 4/2021.

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

LO 8.1 Define *accounting* and its four branches: financial accounting, managerial accounting, auditing, and tax accounting.

LO 8.2 Identify representative accounting questions addressable by business analytics, by branch of accounting.

LO 8.3 Enumerate the sources of accounting data.

LO 8.4 Describe and provide examples of descriptive analytics used in accounting.

LO 8.5 Describe and provide examples of the types of accounting questions addressed by diagnostic analytics.

LO 8.6 Describe and provide examples of the types of accounting questions addressed by predictive analytics.

LO 8.7 Explain how to answer accounting questions with prescriptive analytics.

LO 8.8 Explain how accounting analytics reports results.

LABS FOR CHAPTER 8

LAB 8.1 **Excel:** Descriptive Analytics: Performing Horizontal Analysis

LAB 8.2 **Excel:** Descriptive Analytics: Performing Vertical Analysis

LAB 8.3 **Excel:** Diagnostic Analytics: Using Fuzzy Matching to Look for Fraud

LAB 8.4 **Excel:** Diagnostic Analytics: Estimating Fixed and Variable Costs

LAB 8.5 **Excel:** Predictive Analytics: Forecasting Future Performance of IBM and Netflix

LAB 8.6 **Tableau:** Predictive Analytics: Forecasting Future Performance of IBM and Netflix

LAB 8.7 **Power BI:** Predictive Analytics: Forecasting Future Performance of IBM and Netflix

LAB 8.8 **Excel:** Prescriptive Analytics: Using Goal-Seek Analysis to Determine the Break-Even Point

LO 8.1

Define *accounting* and its four branches: financial accounting, managerial accounting, auditing, and tax accounting.

8.1 THE ROLE OF ACCOUNTING IN BUSINESS

As the financial language of business, accounting provides financial information to decision-makers. **Accounting** is a systematic process of measuring, recording, and communicating the financial information that decision-makers need to make decisions. Accounting also classifies, verifies, and summarizes business transactions in a way that makes the information useful and accessible to decision-makers.

How might managers use accounting? **Adobe** uses accounting to determine how much tax it owes and to help decide whether to pay bonuses to employees. At **Tyson**, managerial accountants track the cost of a pound of chicken to help determine where to cut costs. Individual investors use accounting to decide whether to invest in **Facebook**, **DoorDash**, **Airbnb**, **Google**, or any other publicly traded company. **Apple** uses accounting to determine if and when it should release new products and what features they should include.

Whether you want to talk about business, invest in a business, work for a business, or run your own business—you need accounting and accounting information.

The Four Primary Branches of Accounting

Accounting has four primary branches: financial accounting, managerial accounting, auditing, and tax accounting. Each has a different function and informs decision-making in a different way.

Financial accounting focuses on collecting and analyzing business-performance information for external decision-makers, such as stockholders, banks, investors, and suppliers. Financial accountants provide information about the company's performance, including its financial position and its level of profitability. The primary output of financial accounting is financial statements that are prepared in conformity with accounting rules called **generally accepted accounting principles (GAAP)**. GAAP is a common set of accounting principles, rules, standards, and procedures issued by the standard-setting bodies.[1] GAAP is intended to ensure that all companies' financial statements record and communicate information in the same way, allowing for easy comparison between companies and accurate reporting of accounting data.

Managerial accounting focuses on collecting and analyzing business-performance information for internal decision-makers (management), including the chief executive officer (CEO), chief financial officer (CFO), and **controller** (the person responsible for all of a company's accounting-related activities). Managerial accounting does not have a codified set of accounting rules. Rather, it exists to address management's questions using accounting and non-accounting data.

Auditing investigates accounting and financial records (and their related financial statements) to help determine if they conform with GAAP and other relevant standards or laws. Auditors work to find errors, inaccuracies, and fraud. This audit work can be performed by an independent external accounting firm (such as **KPMG** or **BKD**) or by **internal auditors**. These company employees, who report directly to the board of directors, independently and objectively evaluate the organization's financial and operational business activities, including corporate governance.

Tax accounting focuses on collecting and analyzing transactions to ensure they conform with tax law (as opposed to other standards, such as GAAP used in financial accounting). In addition to focusing their efforts on tax compliance, tax accountants work to help their company or client minimize future taxes through tax planning. Companies want to remain in compliance with tax laws, but not pay more in taxes than what the tax law requires.

[1]There are two primary sources of GAAP. U.S. GAAP is promulgated primarily by the Financial Standards Accounting Board. International GAAP is promulgated primarily by the International Accounting Standards Board.

MINI CASE: Analytics Tools Allow Auditors to Find Unusual Transactions

The large public accounting firms offer a variety of analytical tools to their auditors and their clients. Consider **PwC**'s Halo as an example. This tool allows auditors to interrogate a client's data and identify patterns and relationships within the data in a user-friendly dashboard. By mapping the data, auditors and managers can identify inefficiencies in business processes, discover areas of risk exposure, and correct data-quality issues by drilling down into the individual users, dates, times, and amounts of accounting entries. Halo and other similar tools allow auditors to develop their audit plan by narrowing their focus and audit scope to unusual and infrequent transactions that represent a high risk.

Decision-Making with Data

Which type of business analytics identifies unusual and infrequent issues? Why is identifying this information important to auditors and their clients?

PROGRESS CHECK

1. What is the difference between financial accounting and auditing?
2. What is the difference between the managerial accounting function and the tax accounting function at a company?

8.2 SPECIFYING THE ACCOUNTING QUESTION

LO 8.2

Identify representative accounting questions addressable by business analytics, by branch of accounting.

Specifying the accounting question is the first step of the SOAR model in accounting analytics. Each branch of accounting has different types of questions that are potentially answerable by data. Exhibit 8.1 provides examples of accounting questions according to analytics type (descriptive, diagnostic, predictive, prescriptive) and branch of accounting (financial, managerial, audit, tax). The type of question asked will dictate the remaining steps of the SOAR model, with each question requiring the appropriate data and analytics techniques.

DATA ANALYTICS AND YOUR CAREER

EY is one of the four largest accounting firms in the country. EY considers business analytics so important to its business that it recently offered a free master's degree in business analytics to all of its current employees.[2] A master's degree in analytics focuses on advanced software tools (such as SQL and SAS), programming (such as Python or R), and advanced visualizations (such as Tableau or Power BI). Clearly, analytics is becoming an integral part of not only accounting but also all other business disciplines.

[2]"EY and Hult International Business School Announce New Masters in Business Analytics, Free for All EY People," Available online at https://www.ey.com/en_gl/news/2021/10/ey-and-hult-international-business-school-announce-new-masters-in-business-analytics-free-for-all-ey-people. Accessed 8/4/2022.

Accounting Data Analytics Type and Related Accounting Questions	Financial	Managerial	Audit	Tax
Descriptive: What happened? What is happening?	What is the balance of inventory on hand? What are the descriptive performance statistics (for example, means, medians, maximums, and minimums) for the retail industry? What were the return on asset, asset turnover, and return on sales ratios last year?	What percentage of **United Airlines**' flights departed on time this past month? What was the revenue growth for each division? Which product is the most profitable for the company?	How long have the existing accounts receivable (amounts that our customers owe us) been outstanding? Which sales come from a related party (preexisting business relationship or common interest)?	How much did we pay in federal taxes last year? What effective tax rate did we pay? How big is the difference between our taxable income and net income?
Diagnostic: Why did it happen? What are the causes of past results? Why are the results different than expectations?	Why did SG&A (sales, general, and administrative) expenses decrease as compared to the industry?	Why is revenue up on the West Coast but down in the Midwest? How can the computation of price variance and labor rate variance help determine the root causes of the operational results? Why did the collectability of receivables (amount our customers owe us) fall in the current quarter (as compared to the past)?	Which large transactions had proper management approval, which did not have proper management approval, and why? Why were checks recorded out of sequence?	Why did the average effective tax rate change? Why did overall income tax increase even though net income did not? Why is the difference between taxable income and net income getting bigger?
Predictive: Will it happen in the future? What is the probability something will happen? Can we forecast what will happen?	Can we forecast future revenues/cash flows/ earnings with reasonable accuracy? What is the chance the company will go bankrupt?	Do we extend credit or not to customers based on customer characteristics, such as credit score, payment history, and existing debt? What forecast of future sales should be used to set the budget?	Can we predict if the financial statements might contain errors? Should we accept this company as an audit client?	Can the IRS use predictive techniques to find those who are evading taxes? Can we predict whether the IRS will audit us or not?
Prescriptive: What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints?	Will **Novartis** be more profitable if the FDA approves its new pharmaceutical product?	Should the company lease or buy its headquarters office building? Should the company make its products or outsource to other producers? What level of sales will allow us to break even? How can revenues be maximized (or costs be minimized) if there is a trade war with another country?	How will audit work scheduling change if we acquire a Korean company as a client? If an auditing company has all 12/31 year-end audit clients, how will it organize its audit work in the new year?	Should we move operations from the United States to Ireland to minimize taxes? How will our taxes change if certain tax laws change after the next U.S. Congress legislative section?

Exhibit 8.1
Accounting Data Analytics: The Different Types of Accounting Questions in the Four Main Branches of Accounting

PROGRESS CHECK

3. Give two examples of questions asked in tax accounting that accounting analytics can address and classify the questions as descriptive, diagnostic, predictive, or prescriptive.
4. Give two examples of accounting questions that can be answered with predictive analytics.

8.3 OBTAIN THE DATA: THE SOURCES OF ACCOUNTING DATA

LO 8.3

Enumerate the sources of accounting data.

Because accounting itself produces financial data, there certainly is a lot of accounting data available to address accounting questions. The following sections detail possible sources of financial accounting data, managerial accounting data, auditing data, and tax accounting data (Exhibit 8.2).

	Financial Accounting	Managerial Accounting	Auditing	Tax Accounting
Accounting Data Sources	• General journal and general ledger • Financial Statements • SEC filings (10-K, 10-Q, etc.) • Conference call transcripts • Press releases	• Budget data • Standard cost data • Point-of-sale transaction data • Supply chain data • Customer relationship management (CRM) data • Human resource (HR) data	• Financial statements • General journal, general ledger, and subledgers (accounts)	• General journal and general ledger (detailed financial transactions) • Tax returns • Tax-specific information

Exhibit 8.2
Accounting Data Sources for the Four Main Branches of Accounting

Financial Accounting Data Sources

The term **financial accounting data** refers to the materials prepared by a company to assist decision-makers external to the company, such as stockholders, banks, and suppliers. Common sources of financial accounting data include the general journal, general ledger, financial statements, government filings, conference call transcripts, and press releases.

The General Journal and General Ledger

Accountants record the financial impact of business transactions by making a journal entry in a **general journal**, which is a record of business transactions that have a financial impact on the company. The journal entries in a general journal are key components of the information coming from a company's financial accounting or financial reporting system. Information such as which accounts are debited or credited (as shown in the following example journal entries), the amounts of those entries, who recorded the journal entry, and who authorized or approved the journal entry are all potentially relevant information to the data analyst.

	Debit	Credit
1/10 Notes Receivable	3,500	
Sales Revenue		3,500
Record revenue and receipt of note from sale to Jalynn Thomas.		
1/11 Notes Receivable	4,000	
Accounts Receivable		4,000
Record reclassification from accounts receivable to notes receivable.		
1/13 Shipping Expense	5,000	
Cash		5,000
To record payment of shipping bill.		

The **general ledger** is the record-keeping system used to summarize the individual journal entries. It forms the basis for the financial statements by summarizing the current balance of all asset, liability, equity, revenue, and expense accounts. The general ledger is therefore a good summary-level source of accounting data. Because the general journal and general ledger are internal data sources, the company will likely only share them internally and with internal and external auditors.

Financial Statement Data

The financial statements are an important source of information for a vast cross-section of company stakeholders, including investors, management, and lenders. The **financial statements** are a collection of reports that communicate a company's financial results, condition, health, and cash flows. The four financial statements, which include footnotes that provide additional information about the statements' amounts, are:

1. *Balance Sheet:* A statement that reflects the company's assets (the resources available to the company to run its business), liabilities (amounts owed), and shareholders' equity at a specific point in time.
2. *Income Statement:* A statement of company performance summarizing revenues and expenses to determine how much money the firm earned (or lost) over a period of time
3. *Statement of Cash Flows:* A statement summarizing cash inflows (cash received) and cash outflows (cash paid out) over a period of time
4. *Statement of Stockholders' Equity:* A statement of the increases and decreases in stockholders' equity (ownership).

Of all the financial accounting data, financial statements are the most frequently used by both internal users and external users. The financial statements themselves are found in a number of different locations. The company provides its annual report containing the financial statements on its investor relations website. In addition, the U.S. Securities and Exchanges Commission (SEC) keeps a repository of financial statements and related information on its EDGAR website, explained further later in the chapter. **Calcbench** (https://www.calcbench.com/) and **Yahoo! Finance** are websites that include financial reports along with stock market information. For example, Exhibit 8.3 shows the income statement for **Apple Inc**. as it appears on Yahoo! Finance. The financial statements are an excellent place to start accounting data analysis.

Corporate Securities and Exchange Commission Filings (10-Ks, 10-Qs, 8-Ks)

Publicly traded companies trading on exchanges such as NASDAQ or NYSE are required to submit financial statements and other reports to the **Securities and Exchange Commission (SEC)**. The SEC mandates that accounting information from publicly traded

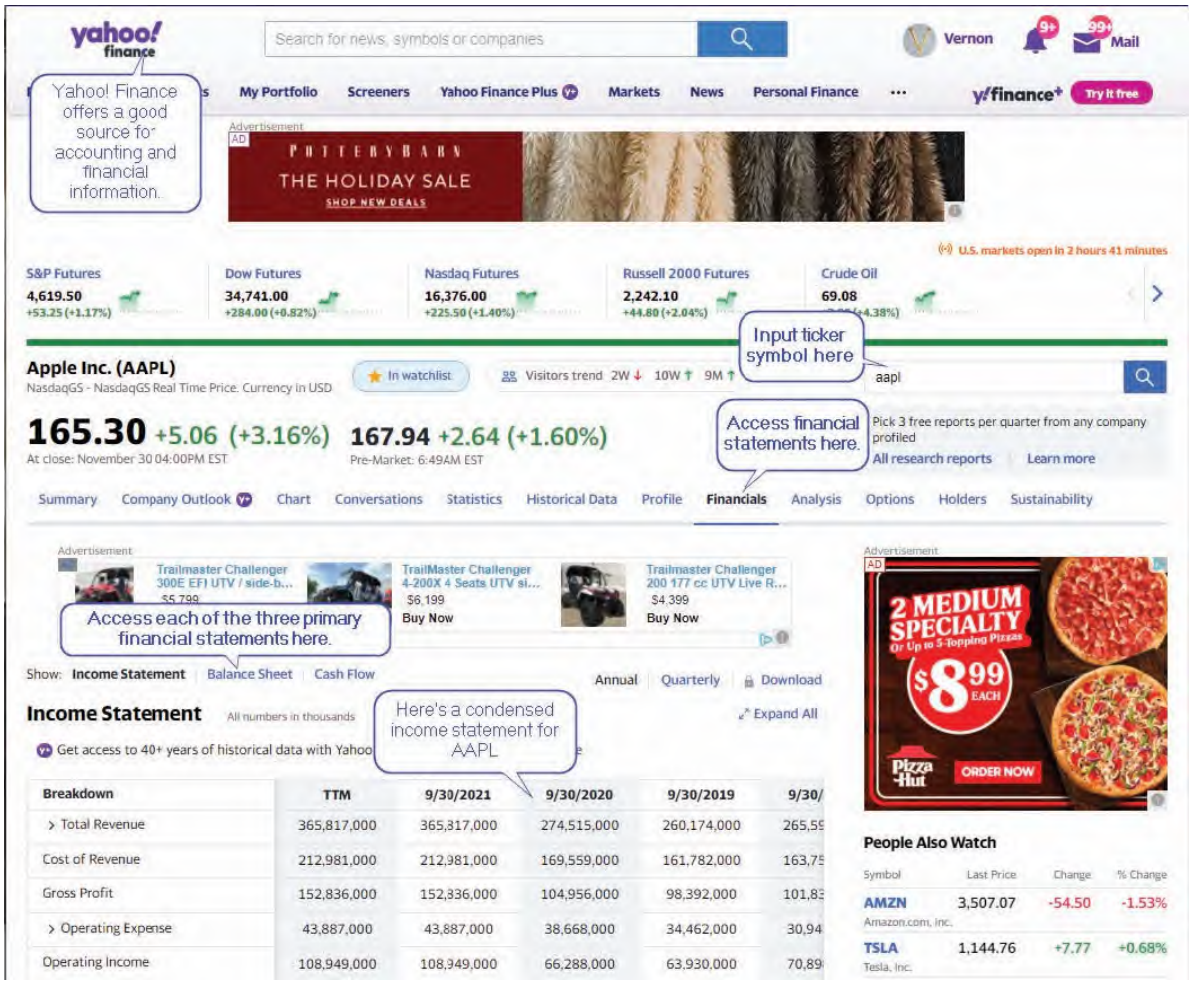


Exhibit 8.3
Income Statement for Apple as shown on Yahoo! Finance

(Source: Yahoo! Finance, accessed December 1, 2021.)

companies be disclosed annually in a **Form 10-K** to ensure that everyday investors have all pertinent information they need to make an investment decision. While this information is provided to all investors, and the SEC is interested in protecting all investors, it is arguably most focused on protecting small investors, typically individuals.

The Form 10-K has five mandatory sections:

- *Business:* An overview of the company's main operations, including the products and services it provides and how it makes money.
- *Risk Factors:* Any significant risks the company faces currently or expects it will, or could, face in the future.
- *Selected Financial Data:* Financial data highlights over the last five years.
- *Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A):* Management's explanation of the company's results from the most recent time period. Unlike most of the rest of Form 10-K, which provides structured

data, the MD&A provides unstructured data that may require additional work to extract and analyze.
- *Financial Statements and Supplemental Data:* The company's audited financial statements, notes to the financial statements, and a report from the auditor explaining the results and scope of the audit.

The related quarterly submission to the SEC required by public companies is **Form 10-Q**. It is similar to Form 10-K in many ways, but it is more abbreviated in its required disclosures. A company typically produces three 10-Q submissions, one for each of the first three quarters of the year. The final quarter is included in the annual 10-K submission.

Another required submission to the SEC is **Form 8-K**, which is used to notify investors of important events or announcements that are not made in the ordinary course of business. Form 8-Ks include information such as the appointment or departure of senior officers, a material or substantial asset acquisition or sale, change in accountants (audit firm), and restatements of financial statements. These reports are used by the SEC, stock analysts, and investors to evaluate a company's performance.

Conference Call Transcripts

An **earnings call** is a conference call between company management (such as the CEO and CFO) and parties external to the company (such as analysts, investors, and the media) to discuss earnings and other financial results that are released immediately preceding the call. Analysts and investors have an opportunity to ask management questions during the call to help them better understand past sales and earnings reports and predict future performance.

Seeking Alpha is one company that offers transcripts of earnings calls; Exhibit 8.4 provides an excerpt. After the call, the transcripts are available for download and analysis. Some financial analysts and accounting researchers perform textual analysis to evaluate the tone and sentiment of management responses to questions to learn about management's general attitude regarding the company's future potential. Transcripts are generally considered unstructured data that require significant analysis to extract tone and sentiment.

Press Releases

Companies often issue **press releases** to make an official statement to the media about a specific matter. Over time, these press releases can be aggregated into a fairly comprehensive view of what is happening at a company, management's approach to business opportunities, and future profitability. Press releases therefore provide unstructured data that are available for analysis.

Managerial Accounting Data Sources

Managerial accounting data help an organization's managers and leaders, such as the CFO and CEO, make decisions. While there are no rules or GAAP standards for managerial accounting data, these data generally meet two criteria:

1. They are useful to internal stakeholders in making decisions.
2. The cost of capturing, measuring, analyzing, and reporting the data must be less than the value derived (such as insights gained) from the analyzed data.

There are many sources of managerial data, including budget data, standard cost data, point-of-sale transaction data, supply chain data, customer relationship management (CRM) data, and human resource (HR) data.

Exhibit 8.4
Example of Earnings Call Transcript

(Source: https://seekingalpha.com/article/4467854-krispy-kreme-doughnuts-inc-dnut-ceo-michael-tattersfield-on-q3-2021-results-earnings-call. Accessed 11/18/2021.)

Krispy Kreme Doughnuts, Inc. (DNUT) CEO Michael Tattersfield on Q3 2021 Results - Earnings Call Transcript

Nov. 10, 2021 9:10 AM ET | **Krispy Kreme, Inc. (DNUT)**

▶ Play Earnings Call

Krispy Kreme, Inc. (NASDAQ:DNUT) Q3 2021 Earnings Conference Call
November 9, 2021 5:00 PM ET

Company Participants

Rob Ballew - VP, IR

Michael Tattersfield - CEO

Josh Charlesworth - CFO and COO

Joey Pruitt - CAO

Operator

[Operator Instructions] Our next question comes from the line of Brian Mullen from Deutsche Bank. Your line is now open.

Brian Mullen

Thank you, just question on the DFD business in the international segment had nice sequential DFD door growth in the quarter again. Could you speak to the ultimate DFD expansion opportunity over the next several years, specifically, across this current group of company on markets? Do you expect additional growth across all the markets? Just any color on the ultimate DFD opportunity in that segment relative to the approximately 2400 doors that you have today? Thank you.

Michael Tattersfield

So again, we reiterated our long term growth in some of those links to the points of access. We see point of access growth of 800 to 1,000 on a yearly basis, half of that being outside of the United States, even within the existing countries that we're in today. This does not even include, as we open up with and leverage our existing base and new partnership countries start to open up. They will start to think about the points of access approach as well. So there is significant growth inside of international both of growing points of access and we continue to see.

Budget Data

A **budget** serves as a financial plan for an upcoming month, quarter, or year and is used to prioritize the organization's needs and allocate resources. Budgets generally start with a prediction of the level of sales the company expects in a forthcoming time period. After determining the estimated level of sales, the company then predicts the level of expenses (such as cost of goods sold, advertising expense, sales expense, and salaries) and the large (capital) expenditures needed to support those sales, such as factory equipment, computers, and office equipment.

As the year progresses and actual outcomes (such as sales and related expenses) become known, it is useful to compare the actual results against the budget. Comparing the actual amounts to the budgeted amounts helps the company determine where results vary from expectations and decide if further adjustments need to be made.

Standard Cost Data

Standard cost is an estimated or expected cost for one unit of production. For example, **Tyson** headquarters in Springdale, Arkansas, processes millions of chickens each week. The company determines the standard cost per pound of chicken and works to understand how changes in feed costs, transportation costs, and electricity costs will affect its overall production costs and ultimately its profits. Knowing the standard cost also helps Tyson determine its level of output (specifically, the number of pounds of chicken to produce) and inputs (including the amount of chicken feed it needs and how it will transport its chicken to retail outlets. Tyson continuously monitors this standard cost of a pound of chicken.

By examining the standard cost data, decision-makers can better assess their options. For example, they may seek out alternate sources of chicken feed if the cost of the existing feed increases too much. Likewise, if the price of electricity rises too high, they can look for other means, such as propane or natural gas, for heating the feed houses. Standard cost data are also an important component in determining the price that should be charged to customers.

Point-of-Sale Transaction Data

A **point-of-sale (POS) system** creates a stream of data from every transaction that occurs. For example, every sale recorded via a cash register at a **Walmart** store is captured and analyzed by Walmart's suppliers. Walmart's POS system, called Retail Link, is an internet-based tool that allows its suppliers to access point-of-sale data. Retail Link provides sales and inventory data—by item, store, and day—to Walmart's suppliers. That is, it contains a record of every sale of every individual item at every Walmart store, every hour of the day for the last two years. Suppliers use these data to predict buying patterns, identify trends, and manage inventory.[3] Walmart's management believes that providing granular, real-time data to suppliers allows them to better predict future demand for their products, helping Walmart ensure that the right products are at the right place at the right time (and hopefully at the right price) to maximize sales and profit.



Quality HD/Shutterstock

Other Corporate Information Systems

The supply chain refers to the process of moving raw materials into production, then from production to distribution, and finally from distribution to the customer. The supply chain often contains many vendors who work at various stages rather than one vendor working from start to finish. A **supply chain management (SCM) system** is an information system that helps manage all of a company's interactions with its suppliers, coordinating the flow of goods and services from original production to final customer delivery. It might include information on active vendors (their contact info, where payment should be made, how much should be paid), the orders made to date (how many orders, when the orders are made), and demand schedules for what component of the final product is needed when.

Recall that a **customer relationship management (CRM) system** is an information system for overseeing all interactions with past, current, and potential customers with the goal of improving relationships. CRM systems contain every detail known about the customer, such as customer contact history, customer order history, and customer payment

[3]https://18knowledge.com/blog/walmarts-retail-link-point-of-sale-data/. Accessed May 2020.



history. Accountants may use data from these systems to get a listing of the company's most profitable customers. When purchasing data about potential customers from third-party data sources, the company is ethically and legally obligated to take steps to ensure customers' privacy.

Companies also maintain a set of data about what is arguably their most valuable asset: their employees. A **human resource management system (HRMS)** is an information system for managing all interactions with past, current, and potential employees, including a listing of potential employee recruits, extent of employee training, employee payroll and compensation, and employee benefits. Accountants are particularly interested in, the mix of salaried and hourly workers, the cost of labor, particularly in a product manufacturing setting.

Auditing Data Sources

An *audit* is an independent investigation of accounting and financial records (and the related financial statements and general journal entries) to determine if they conform with generally accepted accounting principles (GAAP) or other relevant standards or laws. Auditors audit their client's financial accounting systems.

Tax Data Sources

Accountants are interested in how transactions and events affect the amount of tax that is due. As noted earlier in this chapter, they report tax-related data to comply with the tax law and use the data for tax planning purposes to minimize future tax payments.

Tax-Specific Financial Transaction Data

A company's tax information comes primarily from its financial accounting data. In other words, the tax function generally does not generate its own data from transactions, but it does store and maintain certain types of data separately from the financial accounting data. These data types include the following:

1. *Depreciation expense.* Depreciation expense is the allocation of the cost of plant and equipment over time. It is deducted from income to arrive at taxable income. However, because the depreciation method used for tax purposes is almost always different from the depreciation method used for financial accounting purposes, the depreciation for tax purposes needs to be stored in a database as a separate data item.
2. *R&D tax credit.* The Internal Revenue Service gives companies a tax credit for their research and development (R&D) expenses. To receive this credit, companies must document an appropriate level of detail, linking an employee's time directly to a research activity or the use of specific equipment. Companies keep track of all the required data (such as timesheets for all activities involved in R&D, calendars, and project timelines) in case the tax regulators challenge or deny the tax credit.
3. *Sales by tax jurisdiction.* Many jurisdictions (states, cities) require sales tax based on sales in that jurisdiction. This information needs to be tracked and stored, to prove compliance in case of audit by state and local tax authorities.

Tax-Return Data

Taxes are paid to fund the many local, state, and national government programs. Tax returns are good sources of tax data. For example, the IRS and state and local tax authorities use tax returns to assess tax compliance. Using trends from tax returns of similar individuals or companies, the IRS can develop a tax expectation for the year and determine if some taxpayers are evading taxes. In other words, the IRS can use diagnostic analytics to see if the amount of taxes each company pays is different from the amount of taxes the IRS expects it to pay.

PROGRESS CHECK

5. What accounting information is included in the 10-K, the mandatory submission to the Securities and Exchange Commission?
6. How might companies use a point-of-sale system to have the right product available at the right time at the right price and at the right place for its customers?
7. What does the statement of cash flows provide that the income statement does not?

LO 8.4

Describe and provide examples of descriptive analytics used in accounting.

8.4 DESCRIPTIVE ACCOUNTING ANALYTICS

After the question is specified and the accounting data are ready for analysis, analysts are ready for the next step in the SOAR analytics model: ANALYZE THE DATA. This section focuses on descriptive accounting analytics.

Recall that descriptive analytics characterizes, summarizes, and organizes features and properties of the data to facilitate understanding of the results and the underlying data. Descriptive accounting analytics uses statistical and summarization techniques, descriptive statistics, horizontal and vertical analysis, and common-size financial statements, as described below. (We introduced descriptive statistics and horizontal and vertical analysis of financial statements and ratios in Chapter 4. We will briefly review them and then move on to additional descriptive analytics.)

Statistical and Summarization Techniques

Summarization and statistical techniques address the basic questions "What happened?" and "What is happening?" by characterizing, summarizing, and organizing data. These techniques are generally included with spreadsheet software such as Microsoft Excel. Counts, totals, sums, averages, subtotals, minimums, maximums, histograms, pivot tables, standard deviations, and ratio analytics are all descriptive analytics techniques that are available within Excel. Exhibit 8.5 summarizes these techniques.

Statistical Technique (Descriptive Analytics)	Description	Example Accounting Question
Counts	Show how frequently an event occurs	How many times do customers order an add-on for a certain product?
Totals, sums, averages, medians, subtotals	Summarize measures of performance	What is the inventory on hand for a specific product?
Minimums, maximums, standard deviations	Summarize measures showing extreme values to help explain what happened	What was the highest sales refund offered by the customer service department, and who approved it?
Graphs (bar charts), histograms	Use horizontal or vertical bars to show comparisons between categories of data	What has the total accounts receivable balance been over the last three years?
Percentage change from one period to the next using vertical analysis, horizontal analysis, or common-size financial statements	Summarizes changes in financial performance from one period to the next	How much did cost of goods sold as a percentage of revenues change over the past two years?
Ratio analysis such as return on assets, return on sales (profit margin), asset turnover ratios, and debt-to-equity ratios	Calculates important financial performance and financial condition ratios for comparison between companies and over time	Which retail company (**Amazon**, **Walmart**, **Target**, **Kroger**, or **Dillard's**) has the highest return on assets ratio?

Exhibit 8.5
Statistical Techniques Used in Accounting Descriptive Analytics

Descriptive Statistics in Accounting

Descriptive statistics give the basic statistics for accounting data. For example, Exhibit 8.6 shows the mean, median, mode, standard deviation, and other statistics for the total assets, net income, and net sales for 707 retail companies in 2021 in millions of dollars. The statistics were generated by Excel's Data Analysis ToolPak. They allow us to better understand the breadth of the companies' performance. For example, net income ranges from a minimum of −$3,944 million to a maximum of about $21,331 million, and total assets range from a minimum of $3.2 million to a maximum of $321,195 million. We also note that the means are much larger than the median measures of central tendency, which suggests that there are a few really large companies (such as **Walmart** and **Amazon**) and many much smaller retail companies.

	A	B	C	D	E	F
1	*Total Assets*		*Net Income*		*Total Revenue*	
2						
3	Mean	8529.491885	Mean	416.1882	Mean	13614.54
4	Standard Error	1108.258915	Standard Error	64.05468	Standard Error	1875.217
5	Median	1266.911	Median	19.29	Median	1547.502
6	Mode	22360	Mode	1.879	Mode	2483.994
7	Standard Deviation	29468.01894	Standard Deviation	1703.18	Standard Deviation	49861.03
8	Sample Variance	868364140.3	Sample Variance	2900822	Sample Variance	2.49E+09
9	Kurtosis	58.4309544	Kurtosis	56.03447	Kurtosis	63.76864
10	Skewness	7.270650548	Skewness	6.582733	Skewness	7.394555
11	Range	321191.756	Range	25275	Range	556933
12	Minimum	3.244	Minimum	-3944	Minimum	0
13	Maximum	321195	Maximum	21331	Maximum	556933
14	Sum	6030350.763	Sum	294245.1	Sum	9625478
15	Count	707	Count	707	Count	707
16						

Exhibit 8.6
Summary Statistics for 707 Retail Companies Reported in 2021 (Generated Using Microsoft Excel Data Analysis ToolPak)

(Source: Compustat Database of all retail companies with fiscal years ending in 2021.)

Horizontal and Vertical Analysis of Financial Statements

As explained in Chapter 4, horizontal analysis computes the dollar change (Exhibit 8.7, Column D) and percentage change (Exhibit 8.7, Column E) for each line item. In Exhibit 8.7, the percentage change is computed as [new (2021) – old (2020)] / old (2020)] for each line item. So, to compute the percentage change of **Apple**'s sales from products, cell D6 computes the following (in $ thousands):

$$\text{Percentage change (increase or decrease)} = \frac{(\text{New} - \text{Old})}{\text{Old}}$$

$$= \frac{(297{,}392 - 220{,}747) * 100\%}{220{,}747} = 34.7\%$$

The sparklines shown in Column F indicate the trend over the two-year period. **Sparklines** are small visuals that fit within one cell in Excel, visually depicting what happened in the cells next to it.

Also recall from Chapter 4 that vertical analysis expresses financial information in relation to some relevant figure, or base. For example, on an income statement, the relevant base is net sales revenue. The relevant base for a balance-sheet line item is total assets.

Exhibit 8.7
Horizontal Analysis for Apple Inc. Income Statement

(Source: Apple 2021 10-K, Accessed January 2022. From Lab 8.1 Excel)

	A	B	C	D	E	F
1	Apple Inc. (AAPL)					
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)					
3	*Fiscal Period*	Y 2020	Y 2021		Percentage	
4	*Period Start*	9/29/2019	9/27/2020	Change from	Change from	
5	*Period End*	9/26/2020	9/25/2021	2020 to 2021	2020 to 2021	
6	Products	$220,747	$297,392	$76,645	34.7%	
7	Services	$53,768	$68,425	$14,657	27.3%	
8	Net sales	$274,515	$365,817	$91,302	33.3%	
9	Products	$151,286	$192,266	$40,980	27.1%	
10	Services	$18,273	$20,715	$2,442	13.4%	
11	Cost of sales	$169,559	$212,981	$43,422	25.6%	
12	Gross margin	$104,956	$152,836	$47,880	45.6%	
13	**Operating expenses:**					
14	Research and development	$18,752	$21,914	$3,162	16.9%	
15	Selling, general and administrative	$19,916	$21,973	$2,057	10.3%	
16	Total operating expenses	$38,668	$43,887	$5,219	13.5%	
17	Operating income	$66,288	$108,949	$42,661	64.4%	
18	Other income/(expense), net	$803	$258	($545)	-67.9%	
19	Income before provision for income taxes	$67,091	$109,207	$42,116	62.8%	
20	Provision for income taxes	$9,680	$14,527	$4,847	50.1%	
21	Net income	$57,411	$94,680	$37,269	64.9%	

Exhibit 8.8 illustrates vertical analysis for the **Apple Inc**. (Ticker Symbol = AAPL) Income Statement. For each year, each account amount—such as cost of goods sold (cost of sales) or selling, general, and administrative expenses—is expressed as a percentage of net sales for that year. For 2021, the cost of sales of $212,981 is divided by total 2021 net sales of $365,817, which equals 58.22%. For 2020, the number was 61.77%, so it appears that costs decreased from 61.77% of sales to 58.22% of sales (see Row 11).

Also note in Exhibit 8.8, Row 14, that research and development expense decreased from 6.83% of revenues in 2020 to 5.99% of revenues in 2021. Generally, a decrease in research and development expense is viewed as a bad signal for future products, so this decrease could possibly be considered a bad trend. At the same time, selling, general, and administrative expenses decreased from 7.25% to 6.01% of revenues, suggesting that the company is doing a better job of controlling these expenses.

Exhibit 8.8
Vertical Analysis for Apple Inc.

(Source: Apple 2021 10-K, Accessed January 2022. From Lab 8.2 Excel.)

	A	B	C	D	E	F
1	Apple Inc. (aapl)					
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)					
3	*Fiscal Period*	Y 2020	Y 2021			
4	*Period Start*	9/29/2019	9/27/2020	Percent of Sales		
5	*Period End*	9/26/2020	9/25/2021	2020	2021	
6	Products	$220,747	$297,392	80.41%	81.30%	
7	Services	$53,768	$68,425	19.59%	18.70%	
8	Net sales	$274,515	$365,817	100.00%	100.00%	
9	Products	$151,286	$192,266	55.11%	52.56%	
10	Services	$18,273	$20,715	6.66%	5.66%	
11	Cost of sales	$169,559	$212,981	61.77%	58.22%	
12	Gross margin	$104,956	$152,836	38.23%	41.78%	
13	**Operating expenses:**					
14	Research and development	$18,752	$21,914	6.83%	5.99%	
15	Selling, general and administrative	$19,916	$21,973	7.25%	6.01%	
16	Total operating expenses	$38,668	$43,887	14.09%	12.00%	
17	Operating income	$66,288	$108,949	24.15%	29.78%	
18	Other income/(expense), net	$803	$258	0.29%	0.07%	
19	Income before provision for income taxes	$67,091	$109,207	24.44%	29.85%	
20	Provision for income taxes	$9,680	$14,527	3.53%	3.97%	
21	Net income	$57,411	$94,680	20.91%	25.88%	

LAB CONNECTIONS

LAB 8.1 EXCEL and **LAB 8.2 EXCEL** provide an illustration of vertical analysis for **Apple** and **Google**. They also demonstrate the use of sparklines to quickly display what occurred.

Common-Size Financial Statements



Susan Montgomery/ Shutterstock

Common-size financial statements take vertical analysis one step further. In common-size financial statements, each line item is expressed as a percent of net revenues (on the income statement) and percent of total assets (on the balance sheet), and analysts use the numbers to compare different companies. For example, Exhibit 8.9 shows the common-size income statements of **Chipotle** (Ticker = CMG), **Intel** (INTC), **Microsoft** (MSFT), **Disney** (DIS), and **3M** (MMM).

The common-size financial statements help us compare and contrast different companies' financial performance. We can also use them to compare one company over multiple time periods. As the last row of Exhibit 8.9 shows, for each dollar of revenue, Microsoft has the largest profit margin percentage (36.45%) of all the compared companies. Chipotle has the highest cost of goods sold as a percentage of sales (77.53%), while Microsoft has the lowest cost of goods sold as a percentage of sales (24.59%). Intel has the highest depreciation and amortization at 14.92%, suggesting that it uses more depreciable property, plant, and equipment than the other four companies.

Common Size Financial Statements	Chipotle	Intel Corp	Microsoft Corp	Disney	3M
	Company A	Company B	Company C	Company D	Company E
	2021	2021	2021	2021	2021
INCOME STATEMENT					
Sales (Net)	100.00%	100.00%	100.00%	100.00%	100.00%
Cost of Goods Sold	77.53%	29.90%	24.59%	66.94%	47.69%
Selling, General, and Administrative Expense	7.15%	27.24%	27.33%	20.05%	25.70%
Operating Income Before Depreciation	15.32%	42.87%	48.08%	13.01%	26.61%
Depreciation and Amortization	3.37%	14.92%	6.48%	7.58%	5.42%
Operating Income After Depreciation	11.95%	27.94%	41.59%	5.43%	21.19%
Interest Expense		1.26%	1.40%	2.57%	1.35%
Nonoperating Income (Expense)	-0.15%	4.10%	2.10%	1.45%	0.94%
Special Items	-1.03%	-3.32%		-0.51%	-0.38%
Pretax Income	10.77%	27.46%	42.30%	3.80%	20.40%
Income Taxes - Total	2.12%	2.32%	5.85%	0.04%	3.63%
Minority Interest		0.00%	0.00%	0.76%	0.02%
Income Before Extraordinary Items	8.65%	25.14%	36.45%	3.00%	16.75%
Extraordinary Items and Discontinued Operations	0.00%	0.00%	0.00%	-0.04%	0.00%
Net Income (Loss)	8.65%	25.14%	36.45%	2.96%	16.75%

Exhibit 8.9 Common-Size Income Statements for Chipotle (CMG), Intel (INTC), Microsoft (MSFT), Disney (DIS), and 3M (MMM)

(Source: 2021 Financial Statements.)

PROGRESS CHECK

8. How is vertical analysis used to develop common-size financial statements?
9. How does a comparison of one company to a competitor allow analysts to evaluate a company's performance?

LO 8.5

Describe and provide examples of the types of accounting questions addressed by diagnostic analytics.

8.5 DIAGNOSTIC ACCOUNTING ANALYTICS

Diagnostic analytics asks "Why did it happen?" as well as "What are the causes of past results?" and "Why are the results different than expectations?" Recall that diagnostic analytics investigates an underlying cause that cannot be found simply by looking at the descriptive data. Diagnostic questions require various types of analyses, including identifying anomalies and outliers and performing drill-down analysis to determine relations, patterns, and/or linkages among variables. Let's look more closely at each type of diagnostic accounting analytics.

Identifying Anomalies and Outliers

Anomalies and outliers can occur in all branches of accounting. Uncovering and finding possible explanations for anomalies is one component of diagnostic analytics. Exhibit 8.10 details, for each branch of accounting, the accountant's expectation, examples of potential anomalies, and possible explanations.

Branch of Accounting	The Accountant's Expectation	Example of Anomaly	Possible Explanation
Financial accounting	Last year's profit margins are similar to the profit margins of other companies in the same industry.	Profits declined by 30 percent in the current year as compared to past years.	New products sold well, but expenses were not under control.
Managerial accounting	Standard costs set expectations for specific costs the company will incur.	At **Tyson**, the costs of raising chickens have been higher than expected.	The rising fuel costs of hauling chicken and chicken feed have increased the overall costs of raising chickens.
Audit (external)	Generally accepted accounting principles (GAAP) are applied correctly.	**Deloitte** found that its client had made some accounting errors by recording sales in the incorrect accounting period.	Sales were recorded for products not yet shipped.
Audit (internal)	A company adopts rules for making large cash disbursements—two signers are required for checks over $30,000.	An internal auditor for **Arizona State University** finds that 30 checks over the $30,000 threshold had only one signer, violating the internal control policy.	There are only two possible signers for checks, and one was out on family leave for an extended period of time.
Tax	Taxable income generally increases when financial net income increases.	Taxable income decreases at the same time that financial net income increases.	The tax law changed, making some types of income not subject to tax.

Exhibit 8.10
Diagnostic Accounting Analytics: Examples of Outliers

Often the first step in diagnostic analytics is to look for and identify unusual, unexpected results or transactions. Once we have identified the unexpected results and completed an initial review of what might have occurred, we can investigate further to understand exactly why it occurred and whether it is an error, fraud, or just an extreme observation. Exhibit 8.11 provides a summary of the techniques for finding anomalies/outliers, example questions, and the types of tests performed.

Diagnostic Analytics Technique for Finding Anomalies/Outliers	Example Accounting Question	Specific Technique Employed
Fuzzy matching (see Chapter 4)	Why are some addresses of approved vendors similar to the addresses of employees who work at the company?	Compare the addresses of employees and vendors to see if they are similar to determine if there is a possibility of fraudulent transactions.
Sequence checks	Are any checks or purchase orders missing? Do the missing items signify errors or fraud, or can they be explained?	Run a sequence check to verify that all checks and purchase orders are accounted for and investigate discrepancies.
Duplicate transactions	Why are there duplicate transactions in the financial reporting records? Are they fraudulent or just errors?	Check all amounts, dates, and vendors to determine why there are duplicate entries.
Benford's Law (see Chapter 4)	Are any of the refunds given fraudulent?	Does the first digit of refunds given to customers conform with Benford's Law?
Variance analysis (typically performed in management accounting)	Why are the labor-rate and labor-use variance for direct labor at the manufacturing plant unfavorable?	Using standard costs, evaluate the labor use and labor rate to determine why the actual is different from what was predicted.
Reconciliation of cash balance between the bank balance and the cash (general ledger) balance	What reconciling items (outliers or anomalies) are needed to adjust the bank and general ledger cash balance?	Compare transactions from our general ledger cash account to the transactions reported by the bank.
Internal controls testing: Are rules set by management followed?	Why were some large cash disbursements not approved by senior management, and do they represent a violation of a company's internal controls (rules)?	Evaluate internal controls (rules) for cash disbursements and check for appropriate authorization.

Exhibit 8.11
Diagnostic Analytics Techniques, Examples, and Type of Tests for Identifying Anomalies/Outliers

The statistical techniques used to look for and identify unusual, unexpected results or transactions include controls testing, the Excel MATCH() function, conditional formatting (to highlight duplicate or unique values), graphing and charting (outlier analysis), minimums and maximums (to find extreme values), fuzzy matching, and data sorts. Let's now look more closely at two examples of diagnostic analytics in accounting: sequence check and drill-down analysis.

Sequence Check: Identifying Missing Checks/Payments

Sometimes an accounting-related entry that appears out of order might also highlight missing data. A **sequence check** is an analysis to determine if the key number field in the record is in the correct ascending or descending order. Its purpose is to potentially find missing data, such as a missing check number, deposit number, or cash receipt number.

EXPECTATION	ANOMALY
Checks are written in ascending (or descending) order.	Checks are missing or certain check numbers are skipped. Why?

Let's consider this anomaly. Each check has a preprinted number. Because checks are written sequentially, those written out of order require additional analysis. Why might checks be out of sequence? Two possibilities are that a check might be written but voided or the check is missing. Company accountants, internal auditors, and external auditors examine bank statements and canceled checks to evaluate sequence gaps, seeking answers to these questions:

1. Were there any sequence gaps?
2. If there were sequence gaps, why did they occur?
   a. What was the nature of the transaction(s)?
   b. Is each sequence gap due to fraud or error?

Fuzzy Matching

Fuzzy matching is a diagnostic analytics technique used to find potential equivalents when there is a less-than-exact match. Auditors use fuzzy matching to help find potential fraud. For example, suppose that a company employee lives at 4451 N. Rolling Hills Drive in Fayetteville, AR 72703 and a potentially fraudulent refund check was sent to 4451 RollingHills, Fayetteville, AR without a zip code. Fuzzy matching would help the accountant or auditor identify non-exact matches, which may point to fraudulent activity.

EXPECTATION	ANOMALY
Addresses of company vendors are different from the addresses of company employees.	Some addresses of active vendors (where the company sends payment) are similar to employees' addresses. Are the vendors legitimate or fraudulent?
Addresses of customers receiving refunds are different from the addresses of the employee offering the refund.	Some customers who received refunds have addresses that are similar to employees' addresses. Are the customers legitimate or fraudulent?

Company accountants, internal auditors, and external auditors run tests to evaluate fuzzy matches of employee, vendor, and customer names to address these questions:

1. Were there any fuzzy matches of employee, vendor, or customer names?
2. If there were fuzzy matches among employee, vendor, or customer names and addresses, why did they occur?
   a. What was the nature of each transaction?
   b. Is the similarity due to fraud (employees making up fake vendors to send money to themselves), error, or just a coincidence?

A fuzzy lookup add-in tool for Excel is available at https://www.microsoft.com/en-us/download/details.aspx?id=15011. (It is not currently available for Mac computers.)

With fuzzy matching, tolerances can be set to note how close of a match is close enough—50 percent, 70 percent, or 95 percent. Column G of Exhibit 8.12, titled "Similarity," shows the similarity scores for the client and employee addresses. These scores range between 0 and 1. The higher the similarity score, the higher the potential match.

In setting the tolerances for fuzzy matching, carefully consider the likely results. If an exact match (of 100 percent) is required, then many possible matches will likely be missed. However, if the tolerance is set too low (say, requiring only a 20 percent match), there will likely be too many potential match candidates to consider, making it much harder to find the true matches.

	A	B	C	D	E	F	G
1	Client_first_name	Client_last_name	Client_address	Employee_first_name	Employee_last_name	Employee_Address	Similarity
2	Michelle	Patriot	606 Monica Road	Torey	Gellett	835 Monica Road	0.5812
3	Scott	Finder	801 Monica Road	Torey	Gellett	835 Monica Road	0.5812
4	Janae	Huitt	2 Alpine Lane	Janette	Le Blond	2 Apple Lane	0.5490
5	David	Rocks	409 Goodland Hill	Ferd	Moult	409 Goodland Hills	0.9710
6	Lilian	Henryson	973 Judy Circle	Konstance	Illem	973 Joe Circle	0.6333
7	Bonnee	Barson	88495 Loeprich Road	Andree	Clinton	88425 Loeprich Road	0.9667
8	Claudelle	Shirland	95 Ridge Oak Court	Cliff	Maxsted	95 Ridge Oak Courts	0.9842
9	Catherine	Everal	6 Laurel Park	Curr	Pickworth	697 Laurel Park	0.6090
10	Rose	Labrow	972 Badeau Circle	Evangeline	Tizzard	72 Badeau Junction	0.5386
11	Vic	Capaldo	9 Pierstorff Point	Amby	Trott	9 Red Cloud Point	0.5619

Exhibit 8.12
Fuzzy Matching and Similarity Scores (from Lab 8.3 Excel)

SOARing to Success

Trading Off Type I and Type II Errors

If tolerances are set too low, there will be too many possible matches; this is a Type I error. If the tolerances are set too high, requiring almost an exact match, there will be too few possible matches; this is a Type II error (excluding an item when it should not be excluded). Researchers are forever trying to trade off Type I and Type II errors by trying to find the appropriate level to set the tolerance. (See Chapter 3 for a review of Type I and Type II errors.)

LAB CONNECTION

LAB 8.3 EXCEL asks you to use Excel's fuzzy matching feature to evaluate employee, vendor, and customer names and addresses.

Performing Drill-Down Analysis to Determine Relations, Patterns, and Linkages Among Variables

Diagnostic analytics can also uncover additional details in the data by drilling down and summarizing data at different levels to help analysts understand why something happened. With this approach, we look for patterns and linkages in the underlying data set or an external data set to identify potential relationships. Exhibit 8.13 provides a summary of the techniques used in drill-down analysis, example questions, and the types of tests performed. The statistical techniques look for patterns in the underlying data set to identify potential correlations. They include **cross-tabulations** (a method of quantitatively analyzing the relationship between two or more variables, as in pivot tables), data sorts, data manipulation, correlations, and regressions.

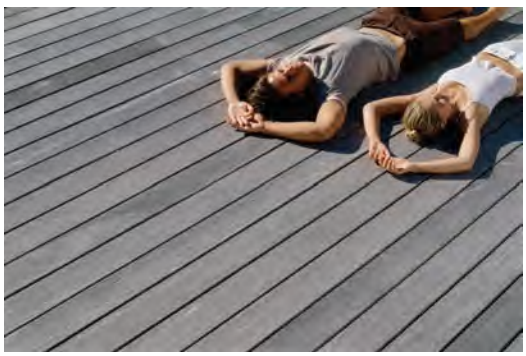
Let's now look more closely at some of these diagnostic analytics techniques.

Use of Regression to Evaluate the Cost Structure of the Company's Production

Understanding its cost structure is important for a company's decision-making. How do company costs depend on the level of production? *Cost behavior analysis* helps a company understand what portion of its production costs are fixed costs and what portion of production costs are variable costs (or mixed costs with part fixed costs and part variable costs).

Diagnostic Analytics Technique for Performing Drill-Down Analysis	Example Accounting Question	Specific Technique Employed
Summarize the data at different levels	Which customers have the highest balance outstanding? Do these customers have invoice amounts that may never be collected?	Sum all of the accounts receivable by customer, and then sort the total accounts receivable balances by customer from highest balance to lowest balance.
Use cross-tabulations (method of quantitatively analyzing the relationship between two or more variables) to view transactions from a different perspective	Why were some transactions approved and recorded on the weekend?	Cross-tabulate using an Excel pivot table to see who approved transactions and the day of the week on which the transactions were approved and recorded.
Identify correlation (the extent to which variables are related to each other)	To what extent are advertising expenses associated with sales revenue?	Run correlation analysis to assess the relationship between advertising expense and sales revenue.
Run a regression (to assess how a specific outcome value is related to variable inputs)	How are R&D expenditures and capital expenditures related to the company's future profitability?	Use regression analysis to assess the relationship between R&D expenditures and net income.
Hypothesis testing	Are sales returns higher during the holiday season, or higher during the non-holiday season?	Statistically compare (using a *t*-test introduced in Chapter 3) the mean sales returns during the holiday season with the mean sales returns during the non-holiday season.

Exhibit 8.13
Diagnostic Analytics Techniques, Examples, and Tests for Performing Drill-Down Analytics to Determine Relations, Patterns, and Linkages Among Variables



Purestock/SuperStock

Knowing the cost behavior can help a company improve job-pricing and job-costing decisions. For this reason, the company may conduct regression analyses to gain a deeper understanding of its cost behavior structure.

Suppose **Trex** is trying to understand its production of wood/plastic composite decking materials (https://www.trex.com). To understand its costs, Trex could run a regression, specifying the dependent variable (outcome) and independent variables (inputs) to view the extent to which its production costs depend on the length in feet of the boards used to produce its wood/plastic composite decking.

- Dependent (outcome) variable: Total production costs (all fixed and variable costs)
- Independent variable: Board feet (number of board feet of decking produced)

The regression model would be set up as:

$$\text{Total production costs} = \beta_0 + \beta_1(\text{board feet})$$

where β_0 is the estimate of Trex's fixed costs and β_1 is the estimate of the variable cost per board foot of composite decking that would be determined in a regression. (β is the Greek letter beta.)

LAB CONNECTION

LAB 8.4 Excel provides step-by-step discussion of and practice with estimating cost behavior at Trex based on costs and the number of board feet produced.

PROGRESS CHECK

10. Why is looking for anomalies an example of diagnostic analytics? How important is the expectation to finding an anomaly? Why are anomalies useful for finding fraud?
11. How does a company working to understand its own cost structure help it determine the strategic options it might have in the future?

8.6 PREDICTIVE ACCOUNTING ANALYTICS

LO 8.6

Describe and provide examples of the types of accounting questions addressed by predictive analytics.

Predictive analytics addresses the question "Will it happen in the future?" along with "What is the probability something will happen?" or "Can we forecast what will happen?" The idea is to predict future performance based on statistical techniques—for example, predicting sales to produce just enough product to meet customer demand without overproducing.

Predictive analytics is composed of three techniques that help companies determine if something will happen in the future: classification, regression, and time series analysis. Exhibit 8.14 summarizes these techniques along with example questions and specific techniques.

Predictive Analytics Technique	Example Accounting Question	Specific Technique Employed
Classification: Separates or classifies a sample (or population) into two or more groups or classes	Which audit clients are at greater risk of misstating financial statements?	Run classification analysis to predict which audit clients are likely to misstate their financial statements.
Regression: Predicts a specific output value based on variable inputs	Does our spending on social media and print media predict the company's level of sales?	Run regression analysis to predict the level of sales based on social media and print media expenditures.
Time series analysis: Predicts future values based on past values of the same variable	Can we predict next year's fuel expenditures based on past expenditures for fuel for the company's truck fleet?	Use time series analysis incorporating past values of fuel expenditures to predict next year's fuel expenditures.

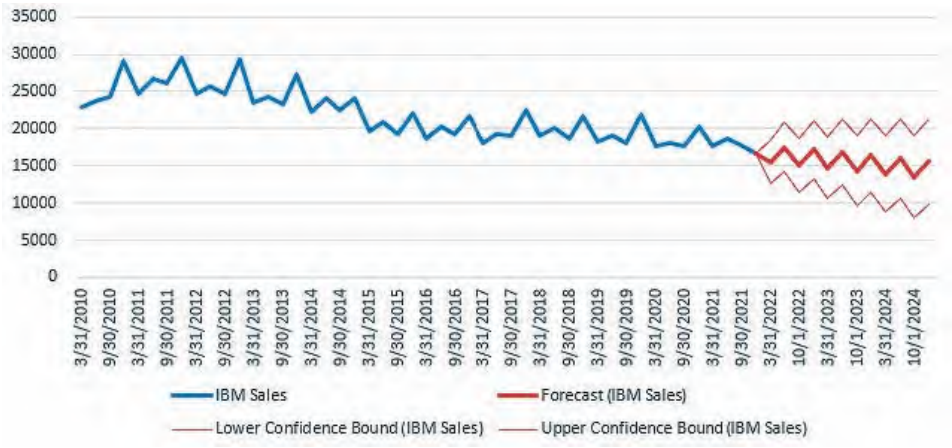
Exhibit 8.14
Predictive Analytics Techniques, Examples, and Tests

Using Time Series Analysis to Predict Sales

Exhibit 8.15 provides an example of time series quarterly analysis for a sales forecast for **IBM** made using Excel's Forecast Sheet. We note some seasonality, or variation by quarter. That is, some quarters consistently have higher or lower sales than other quarters. Also, note the forecast of future sales beginning in 2022 and ending in 2024, as shown by the boldfaced red line. There is also a lower and an upper confidence bound, equal to ± two standard deviations around the predicted level of sales, representing the worst- and best-case scenarios. There is a 95 percent chance that the future actual sales will be in the range between the lower confidence bound and the upper confidence bound.

In general, sales is associated with more persistence from the past to the future than does net income (not shown in Exhibit 8.15), suggesting that time series analysis is more accurate at predicting sales than it is at predicting net income.

Exhibit 8.15
Example of Time Series Analysis Forecasting Quarterly Sales for IBM (Using Excel Forecast Sheet)



LAB CONNECTIONS

In **LAB 8.5 Excel**, you will use Excel's Forecast Sheet to complete a time series analysis similar to that shown in Exhibit 8.15. Specifically, you will use recent sales and earnings performance to forecast future sales and future income for **IBM** and **Netflix**.

PROGRESS CHECK

12. Why are investors interested in a forecast of a company's sales, earnings, and cash flows? Why do analysts need to use predictive analytics to make these forecasts?
13. Why would a company be interested in a forecast of its sales?

LO 8.7

Explain how to answer accounting questions with prescriptive analytics.

8.7 PRESCRIPTIVE ACCOUNTING ANALYTICS

Prescriptive analytics answers the questions "What should we do, based on what we expect will happen?" and "How do we optimize our performance based on potential constraints?" Prescriptive accounting analytics uses five techniques to help us determine what we should do based on what we expect will happen:

1. What-if scenario analysis, including marginal/incremental analysis
2. Cash flow analysis
3. Scenario analysis
4. Goal-seek analysis
5. Sensitivity analysis

Exhibit 8.16 summarizes these techniques, along with example questions and the tests used.

In the following sections, we offer specific examples of goal-seek analysis and sensitivity analysis.

Prescriptive Analytics Technique	Example Accounting Question	Specific Technique Employed
Marginal/incremental analysis: What-if analysis that determines the change in profit associated with the cost or benefit of the next, or marginal, unit	Should the **University of Connecticut** host its own IT infrastructure (web services, email, etc.) or outsource to **Amazon Web Services (AWS)?**	Estimate the financial implications of two scenarios: host own infrastructure or outsource to Amazon.
Cash flow analysis: Evaluates future cash flows for potential investments or expenditures, typically using net present value (NPV) or internal rate of return (IRR)	In which asset should **Boeing** invest: new manufacturing equipment in Seattle, Washington, or a new computer system in Shanghai, China?	Estimate the amounts, timing, and risk associated with the predicted cash flows from the two possible investments.
Scenario analysis: What-if analysis that predicts how changing multiple inputs at the same time will affect an outcome variable	How does **Netflix** respond to various scenarios of tax-law changes?	Analyze all of the potential scenarios to arrive at an overall potential impact of the potential tax law change.
Goal-seek analysis: What-if analysis that determines the required input value(s) to get a desired output (that is, backward analysis)	At **Canoo** (a new electric vehicle manufacturer), what level of sales is needed to break even?	Perform break-even analysis to figure out the needed level of sales to break even.
Sensitivity analysis: What-if analysis that predicts how changing one input will affect an outcome variable	What happens to **Wells Fargo**'s profits if interest rates increase from 3 percent to 4 percent or decrease to 2 percent?	Test to see how sensitive Wells Fargo's profits are to interest-rate changes based on our predictive models.

Exhibit 8.16
Prescriptive Analytics Techniques, Examples, and Specific Tests

Goal-Seek Analysis

Recall that *goal-seek analysis* is a form of what-if analysis that tells us the inputs required to reach a desired outcome, output, or result. Many companies use goal-seek analysis to compute break-even sales—that is, the amount that the company needs to achieve in sales to break even or have exactly $0 profits. At the break-even point, total revenues equal total expenses. Goal-seek analysis works backward: The analyst identifies the desired outcome, and the application determines appropriate input to achieve that outcome.

Exhibit 8.17 is a screenshot from Lab 8.6 Excel showing the use of the Goal Seek function in Excel. Specifically, it shows the amount of January unit sales (cell B3) required to make net income exactly equal to zero (cell B9) or to break even.

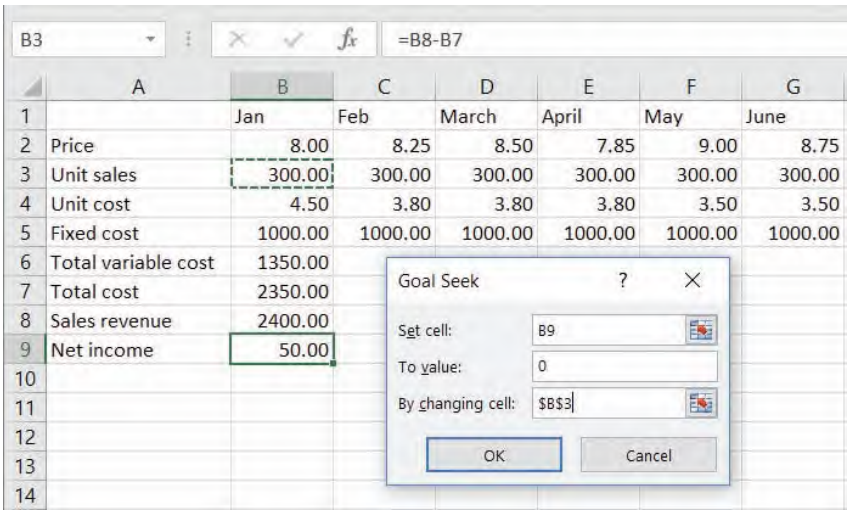


Exhibit 8.17
Goal-Seek Analysis Showing Level of Sales Needed to Break Even (from Lab 8.6)

LAB CONNECTION

LAB 8.6 EXCEL provides step-by-step practice with using goal-seek analysis to calculate the break-even level of sales.

Sensitivity Analysis

Recall that *sensitivity analysis* is a type of what-if analysis that examines the effect of changing one input on an outcome variable. Here are a few examples of how companies might use sensitivity analysis.

- If gas prices increase by 10, 15, or 20 percent, what will be the impact on **Southwest Airlines**' net income? Because airline fuel is a critical expense for any airline, management will want to carefully plan around the potential changes in gas prices from year to year.
- **Ford Motor Company** imports steel from China to use in the production of its cars. If exchange rates between the Chinese yuan and the U.S. dollar increase 10, 20, or 30 percent, how will that change affect the cost of steel as a key input to car prices for Ford? If unfavorable exchange rates greatly increase the price of steel, Ford may decide to purchase steel futures to lock in the price of steel and thus not be susceptible to changing exchange rates and steel prices.

To perform a sensitivity analysis, management accountants could use a table similar to Exhibit 8.18 to evaluate the company's profitability based on two important inputs: (1) total fixed costs and (2) the sales price of the product. The table holds constant other important variables, such as per-unit variable costs ($20) and total units sold (300).

As the exhibit shows, the analysis varies the sales price input from $35 to $53 (as shown in the columns) and the fixed costs input from $500 to $3,000 (as shown in the rows). The colored columns are the total profits based on the changing input parameters. For example, if the sales price is $53 and fixed costs are $3,000, then $6,900 in profits can be expected (assuming per-unit variable costs of $20 and 300 units sold).

		Sales Price									
		$ 35.00	$ 37.00	$ 39.00	$ 41.00	$ 43.00	$ 45.00	$ 47.00	$ 49.00	$ 51.00	$ 53.00
Fixed Costs	500	4000.00	4600.00	5200.00	5800.00	6400.00	7000.00	7600.00	8200.00	8800.00	9400.00
	750	3750.00	4350.00	4950.00	5550.00	6150.00	6750.00	7350.00	7950.00	8550.00	9150.00
	1,000	3500.00	4100.00	4700.00	5300.00	5900.00	6500.00	7100.00	7700.00	8300.00	8900.00
	1,250	3250.00	3850.00	4450.00	5050.00	5650.00	6250.00	6850.00	7450.00	8050.00	8650.00
	1,500	3000.00	3600.00	4200.00	4800.00	5400.00	6000.00	6600.00	7200.00	7800.00	8400.00
	1,750	2750.00	3350.00	3950.00	4550.00	5150.00	5750.00	6350.00	6950.00	7550.00	8150.00
	2,000	2500.00	3100.00	3700.00	4300.00	4900.00	5500.00	6100.00	6700.00	7300.00	7900.00
	2,250	2250.00	2850.00	3450.00	4050.00	4650.00	5250.00	5850.00	6450.00	7050.00	7650.00
	2,500	2000.00	2600.00	3200.00	3800.00	4400.00	5000.00	5600.00	6200.00	6800.00	7400.00
	2,750	1750.00	2350.00	2950.00	3550.00	4150.00	4750.00	5350.00	5950.00	6550.00	7150.00
	3,000	1500.00	2100.00	2700.00	3300.00	3900.00	4500.00	5100.00	5700.00	6300.00	6900.00

Exhibit 8.18
Sensitivity Analysis: An Illustration

PROGRESS CHECK

14. What is sensitivity analysis, and why is it considered a type of prescriptive analytics?
15. What is the desired output in a break-even analysis? What is the company trying to determine?

MINI CASE: Lebron James, What-If Analysis, and Taxes



Li Ying/Xinhua/Alamy Live News

Knowing the tax liability involved with a move to a new jurisdiction is important for corporations and individuals alike. For example, a tax accountant might have advised LeBron James not to sign with the Los Angeles Lakers in the summer of 2018 because that decision was expected to cost him $21 million more in state income taxes because California has higher taxes than Ohio. What-if analysis could have helped frame LeBron's decision from a tax-planning perspective. We have to assume that LeBron believed that the benefits of moving to L.A. were worth more than the extra $21 million in state taxes that he now pays.

What-if analysis has wide application when a company is trying to determine the likely effects of potential new legislation, evaluate a merger possibility, decide whether to shift its product mix, or determine whether to set up operations in a new low-tax (or high-tax) jurisdiction. For example, **Tesla** recently applied tax-planning concepts when considering the tax incentives available for locating its Cybertruck factory. Tesla ultimately decided to build its factory in Texas.

Decision-Making with Data

What might Tesla do if Ireland offers tremendous tax incentives to move a portion of Tesla's operations there? How might such a move affect Tesla's profits? How might analysts use what-if analysis to determine the best course of action?

8.8 REPORT THE RESULTS

LO 8.8

Explain how accounting analytics reports results.

As noted in Chapter 6, there are many ways to report the results. Tables and data visualizations are very useful in summarizing findings for decision-makers.

Tables Are Sometimes a More Effective Way to Show Results

Exhibit 8.19 provides an example of how vertical analysis, horizontal analysis, and common-size financials are shown in **J.B. Hunt**'s management discussion and analysis section of its 2020 annual report/10-K.

In many cases, visualizations more directly convey the results of data analysis than a data table. But in this case, as Exhibit 8.19 shows, a table is the best way to show the combined

horizontal and vertical analyses and common-size financial statements. The first three columns show the numbers for three specific years: 2019, 2020, and 2021. The last two columns show the percentage change between 2019 and 2020, and between 2020 and 2021. No other type of visualization would show these three types of analyses with so much detail.

Exhibit 8.19
Vertical and Horizontal Analysis and Common-Size Financial Statements in J.B. Hunt's 2021 Annual Report/10-K

In what way does this exhibit provide a horizontal analysis? In what way does it provide a vertical analysis?

(Source: J.B. Hunt 2020 10-K/Annual Report, https://investor.jbhunt.com/financials/annual-reports/default.aspx. Accessed 8/4/2022.)

RESULTS OF OPERATIONS

The following table sets forth items in our Consolidated Statements of Earnings as a percentage of operating revenues and the percentage increase or decrease of those items compared with the prior year.

	Percentage of Operating Revenues			Percentage Change Between Years	
	2021	2020	2019	2021 vs. 2020	2020 vs. 2019
Operating revenues	100.0%	100.0%	100.0%	26.3%	5.1%
Operating expenses:					
Rents and purchased transportation	53.0	51.4	49.4	30.2	9.4
Salaries, wages and employee benefits	22.7	24.4	23.7	17.6	8.3
Depreciation and amortization	4.6	5.5	5.4	5.6	5.7
Fuel and fuel taxes	4.4	3.7	5.1	48.4	(22.8)
Operating supplies and expenses	3.0	3.5	3.6	10.5	0.4
General and administrative expenses, net of asset dispositions	1.5	1.8	2.1	8.6	(6.2)
Insurance and claims	1.4	1.4	1.7	22.7	(14.5)
Operating taxes and licenses	0.5	0.6	0.6	9.4	(1.8)
Communication and utilities	0.3	0.3	0.4	4.0	(3.7)
Total operating expenses	91.4	92.6	92.0	24.6	5.8
Operating income	8.6	7.4	8.0	46.6	(2.8)
Net interest expense	0.4	0.5	0.6	(2.8)	(11.0)
Earnings before income taxes	8.2	6.9	7.4	50.1	(2.2)
Income taxes	1.9	1.6	1.8	49.4	(2.8)
Net earnings	6.3%	5.3%	5.6%	50.3%	(2.0)%



DATA VISUALIZATION

Using Visualizations to Highlight Anomalies

As we've seen, data analysts usually seek to highlight and explain anomalies. To show an anomaly clearly, the analyst needs to show the actual performance and how it departs from the expectation. For example, an analyst looking for potential fraud could use the heat map in Exhibit 8.20 to evaluate relationships between accounts payable clerks (rows) and vendors/suppliers (columns). According to Exhibit 8.20, AP Clerk AP0060 processes over 90% of the payments to Peters Co., which could mean there is an inappropriate relationship between the vendor and the AP clerk. The heat map also details other potential inappropriate relationships that may require further investigation.

Sum of Invoice Amount / Column Labels														
Row Labels	Bobs Inc.	Connie Ent.	Dizzy Ind.	Mayday LLC	North Ind.	Perrys Co.	Peters Co.	Richardson Corp	SEC Inc.	Tree Co	Twos Inc.	XNA Inc.	Young Co.	Yuxin Corp.
AP0010	13.57%	10.51%	4.26%	14.02%	9.31%	10.88%	5.71%	9.24%	6.12%	9.35%	7.47%	5.72%	10.64%	6.11%
AP0020	7.97%	10.89%	7.43%	7.73%	2.94%	9.21%	0.00%	15.17%	14.68%	14.41%	13.35%	12.33%	20.43%	1.73%
AP0030	13.30%	3.52%	29.73%	8.34%	6.89%	13.99%	0.00%	5.61%	15.77%	8.94%	12.74%	9.61%	9.24%	9.86%
AP0040	2.68%	6.23%	14.30%	17.17%	14.53%	16.93%	0.00%	17.36%	24.62%	10.62%	23.38%	13.58%	17.59%	17.25%
AP0050	17.47%	12.67%	15.54%	10.32%	3.94%	4.33%	4.03%	17.89%	3.17%	14.68%	8.14%	16.84%	8.66%	4.77%
AP0060	6.92%	10.45%	8.84%	13.92%	8.94%	11.90%	90.26%	7.46%	12.98%	20.11%	16.00%	1.29%	14.53%	17.07%
AP0070	12.15%	12.98%	8.74%	10.67%	15.43%	7.95%	0.00%	12.79%	10.34%	6.28%	8.87%	15.78%	2.44%	6.34%
AP0080	13.46%	15.83%	3.72%	5.86%	13.50%	13.74%	0.00%	11.24%	7.61%	4.63%	5.39%	22.21%	12.76%	13.70%
AP0090	12.47%	16.91%	7.45%	11.95%	24.52%	11.08%	0.00%	3.25%	4.69%	10.99%	4.67%	2.64%	3.70%	23.17%

Exhibit 8.20
Using a Heat Map to Find Potential

Using a Graph to Show Break-Even Sales

DATA VISUALIZATION

Exhibit 8.21 is an effective way to show the break-even level of sales. This simple graph details the size of the fixed costs ($1,000,000 in this example), the total cost (which includes both fixed costs and variable costs), and the break-even point where total revenue equals total cost. This graph also helps management determine the level of sales required to reach other levels of profit.

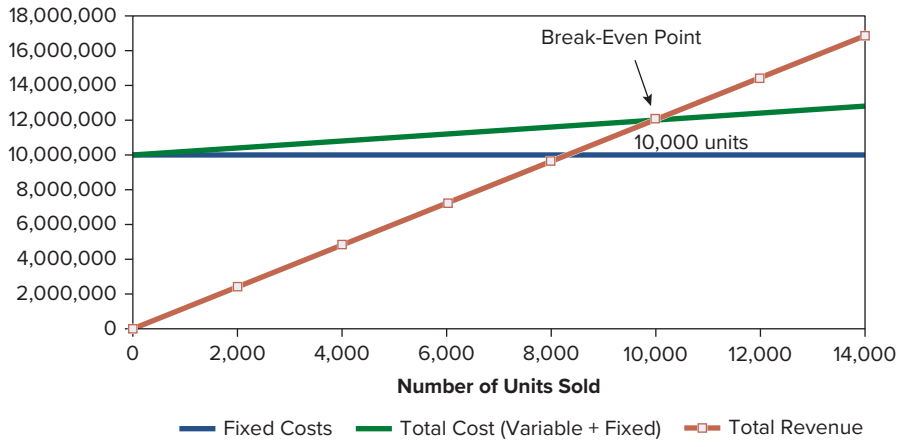


Exhibit 8.21
Break-Even Chart Visualization

PROGRESS CHECK

16. Why are tables sometimes a better way to report results than graphs? What benefits do tables provide?
17. Exhibit 8.19 presents a vertical analysis for J.B. Hunt. Why would J.B. Hunt include this visualization in its Form 10-K for shareholders?

SUMMARY

Accounting analytics is used to address questions in the four main branches of accounting: financial accounting, managerial accounting, auditing, and tax accounting. We discussed the accounting data sources potentially useful in addressing accounting questions. We also discussed the different types of accounting analytics that can be performed, including descriptive, diagnostic, predictive, and prescriptive analytics. The chapter concluded with examples of reporting with tables and visualizations.

Key Terms

accounting (542) A systematic process of measuring, recording, and communicating the information that decision-makers need to make decisions.

auditing (542) Branch of accounting that investigates accounting and financial records (and their related financial statements) to help determine if they conform with generally accepted accounting principles (GAAP) and other relevant standards or laws.

budget (549) A financial plan used to prioritize an organization's needs for an upcoming time period.

common-size financial statements (555) Financial statements in which each line item is expressed as a percent of net revenues (on the income statement) and percent of total assets (on the balance sheet); used to compare different companies.

controller (542) The person responsible for all of a company's accounting-related activities.

cross-tabulation (559) A method to quantitatively analyze the relationship between multiple variables, as in pivot tables.

customer relationship management (CRM) system (550) An information system for managing all interactions with past, current, and potential customers with the goal of improving relationships.

earnings call (548) A conference call between company management (such as the CEO and CFO) and parties external to the company (such as analysts, investors, and the media) to discuss earnings and other financial results that are released immediately preceding the call.

financial accounting (542) Branch of accounting that focuses on collecting and analyzing business performance information for external decision-makers.

financial accounting data (545) Data prepared by a company to assist decision-makers external to the company, such as stockholders, banks, and suppliers.

financial statements (546) A collection of reports that communicate a company's financial results, condition, health, and cash flows.

Form 8-K (548) A required submission to the Securities and Exchange Commission that is used to notify investors of important events or announcements.

Form 10-K (547) A required annual submission to the Securities and Exchange Commission reporting a company's financial performance.

Form 10-Q (548) A required quarterly (every 3 months) submission to the Securities and Exchange Commission reporting a company's financial performance.

fuzzy matching (558) A technique used to find potential matches when there is a less-than-exact match.

general journal (545) A record of business transactions that have a financial impact on the company.

general ledger (546) Record-keeping system used to summarize the individual journal entries that form the basis for the financial statements.

generally accepted accounting principles (GAAP) (542) A common set of accounting principles, rules, standards, and procedures issued by the standard-setting bodies.

human resource management system (HRMS) (551) An information system for managing all interactions with past, current, and potential employees.

internal auditor (542) Company employee who independently and objectively evaluates financial and operational business activities, including corporate governance

managerial accounting (542) Branch of accounting that focuses on collecting and analyzing business-performance information for internal decision-makers (management).

point-of-sale (POS) system (550) A computer system that records all transactions at the point of sale.

press release (548) An official statement to the media from a company about a specific matter.

Securities and Exchange Commission (SEC) (546) U.S. government agency charged with protecting investors by requiring full and truthful financial disclosure by companies.

sequence check (557) A diagnostic analytics test to see if the key number field in a record is in the correct ascending or descending order.

sparklines (553) Small visuals that fit within one cell in Excel visually depicting what happened in the cells next to it.

standard cost (550) Estimated or expected cost for one unit of production.

supply chain management (SCM) system (550) An information system that helps manage all of a company's interactions with its suppliers, coordinating the flow of goods and services from original production to final customer delivery.

tax accounting (542) Branch of accounting that focuses on collecting and analyzing transactions to ensure they conform with tax law and minimizing future taxes through tax planning.

ANSWERS TO PROGRESS CHECKS

1. While financial accounting focuses on collecting and analyzing business-performance information, auditing works to make sure that financial accounting is done correctly (in conformity with GAAP and other relevant standards or laws).
2. Managerial accountants collect and analyze business-performance information for management. Tax accountants ensure that transactions comply with tax law. They also help their company or client minimize taxes through tax planning.
3. Potential questions addressable with accounting analytics include "How much did we pay in taxes last year?" (descriptive analytics) and "Should we move company operations to a different country with a lower income tax rate?" (prescriptive analytics).
4. Potential questions addressable with predictive analytics include "What is the forecasted level of net income next period?" and "Can the IRS predict who is evading taxes?"
5. The SEC Form 10-K contains five required sections: Business, Risk Factors, Selected Financial Data, Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A), and Financial Statements and Supplemental Data. Each of these required sections has accounting data useful for accounting analytics.
6. Point-of-sale data inform analysts what customers are buying as well as when and where. This information reveals the level of demand for each product at each location at the relevant price.
7. While the statement of cash flows reports the amount of cash received (cash inflows) and the cash paid out (cash outflows) over a period of time, the income statement reports the amount of revenue earned and expenses incurred over a period of time.
8. Common-size financial statements are vertical analyses for several different companies shown side by side. Common-size financial statements help us compare and contrast the financial performance of different companies. Common-size statements can also be prepared for one company over multiple time periods.
9. Similar companies face similar economic factors, similar financing factors, and similar competition. By comparing one company to another, analysts can evaluate which company is outperforming the other.
10. One aim of diagnostic analytics is to find anomalies and outliers. Because fraud doesn't happen all the time, the expectations may be that there is no fraud and that transactions are entered in the accounting system fully and completely. Certain transactions may occur that are different than expected, and diagnostic analytics may be able to find these transactions. Once anomalies are found, the auditor will investigate further to determine whether fraud has occurred or not.
11. Companies often compete on quality, cost, or value. If a company can understand its costs, it can better decide how to compete in the marketplace and make profits.
12. Predictive analytics forecast what we expect to happen. Investors are interested in forecasting earnings to decide whether they will invest in the company's equity or debt.

Analysts predict future sales, earnings, and cash flows using time series analysis and other predictive analytics techniques.

13. A company wants to fulfill customer demand. It will use forecasted sales to make sure it has the appropriate resources (such as company employees, supplies, and parts) to meet that customer demand.
14. Sensitivity analysis is a type of what-if analysis that predicts how changing one input will affect an outcome variable. It is a prescriptive analytics technique because it helps decision-makers decide what to do based on what they expect will happen.
15. The desired output is a net income of exactly zero. The company is trying to determine the level of sales that will exactly cover the company's expenses.
16. Tables may be the most efficient way to show the numerous details that a graph or other visualization may not be able to provide.
17. The vertical analysis offers an effective means for the company to communicate its results from one year to the next. It allows J. B. Hunt to display its performance relative to sales.



Multiple-Choice Questions

1. (LO 8.4) A graph displaying net income over the past four years results from which type of analytics?
   a. Descriptive analytics
   b. Diagnostic analytics
   c. Predictive analytics
   d. Prescriptive analytics
2. (LO 8.1, 8.2) Budget data are an example of which kind of accounting data?
   a. Financial accounting
   b. Managerial accounting
   c. Auditing
   d. Tax accounting
3. (LO 8.3) Which form is a required annual submission to the Securities and Exchange Commission?
   a. 10-Q
   b. 10-K
   c. 8-K
   d. Press release
4. (LO 8.6) What type of analytics addresses the question of whether an audit firm should accept or reject an audit client?
   a. Descriptive analytics
   b. Diagnostic analytics
   c. Predictive analytics
   d. Prescriptive analytics
5. (LO 8.5) Assessing a company's cost structure is an example of which type of analytics?
   a. Descriptive analytics
   b. Diagnostic analytics

c. Predictive analytics
d. Prescriptive analytics

6. (LO 8.6) What prescriptive analytics technique is used to determine the break-even level of sales?
   a. Sensitivity analysis
   b. Scenario analysis
   c. Cash flow analysis
   d. Goal-seek analysis
7. (LO 8.6) Which type of analytics includes classification, regression, and time series analysis techniques?
   a. Descriptive analytics
   b. Diagnostic analytics
   c. Predictive analytics
   d. Prescriptive analytics
8. (LO 8.8) A heat map is an example of which type of visualization?
   a. Conditional highlighting
   b. Conditional formatting
   c. A temperature gauge
   d. Horizontal analysis
9. (LO 8.6) Which predictive analytics technique is generally used to forecast future sales, earnings, and cash flows?
   a. Time series analysis
   b. Classification
   c. Regression
   d. Goal-seek analysis
10. (LO 8.4) Which analytics technique evaluates the percentage change in performance from one period to the next?
    a. Vertical
    b. What-if
    c. Classification
    d. Horizontal

Discussion Questions

1. (LO 8.1) Describe the role of tax accounting.
2. (LO 8.1) Managerial accounting does not have any rules and standards. How do managerial accountants know which information to analyze?
3. (LO 8.1) Describe questions that external and internal auditors can use business analytics to answer.
4. (LO 8.2) What are some examples of diagnostic analytics for those in financial accounting? Managerial accounting?
5. (LO 8.3) Describe sources of financial accounting data helpful in addressing questions associated with descriptive analytics.
6. (LO 8.4) How does vertical analysis serve as a descriptive analytics technique to describe what is happening?
7. (LO 8.5) Which type of analytics looks for anomalies and outliers? Explain.
8. (LO 8.6) What is the difference between the predictive analytics techniques of classification and regression?

9. (LO 8.7) Which type of analytics will analysts use to optimize company performance based on expected changes in tax law? Explain.
10. (LO 8.7) Which type of analytics will a company use to answer the question of whether it should rent or own its headquarters office building? Explain.
11. (LO 8.8) How does conditional formatting help analysts find anomalies?

connect

Brief Exercises

1. (LO 8.1) Match the branch of accounting (financial, managerial, audit, or tax) to each example accounting question.

Example Accounting Question	**Branch of Accounting** (Financial, Managerial, Audit, or Tax)
Are there errors in the financial statements provided by our clients?	
What is the overall financial performance for the automotive industry?	
What is the cost of production at our new factory?	
How big is the difference between taxable income and net income?	
Should we move our call center operations to the Bahamas to minimize taxes?	
What would be the impact of increased wages for the employees at our call center?	
Do the accounting records reflect all transactions that took place in the month of April?	

2. (LO 8.4, 8.5, 8.6, 8.7) Match the analytics type (descriptive, diagnostic, predictive, or prescriptive) to each example accounting question.

Example Accounting Question	**Analytics Type** (Descriptive, Diagnostic, Predictive, or Prescriptive)
Should we extend credit to a supplier?	
Will the company be profitable if it receives regulatory approval to sell in Europe?	
Why is revenue up for our company but down for the industry as a whole?	
Has net income increased over the past five years?	
Why were actual production costs different from the amount budgeted?	
What level of sales will allow us to break even?	
Were checks recorded out of sequence? Are any checks missing?	

3. (LO 8.3) Match the accounting data sources to the branch of accounting (financial, managerial, audit, or tax).

Accounting Data Source	Branch of Accounting (Financial, Managerial, Audit, or Tax)
SEC filings (10-K, 10-Q)	
Point-of-sale information	
Press releases	
Budget data	
Detailed depreciation records or R&D information to qualify for IRS deductions	
Conference calls	
Tax returns	

4. (LO 8.2) Categorize each data analytics question below as one of the following question types. Answers may be used more than once.
   a. What happened? What is happening?
   b. Why did it happen? What are the causes of past results?
   c. Will it happen in the future? What is the probability something will happen? Can we forecast what will happen?
   d. What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints?

Data Analytics Question	Question Type
How much did we pay in federal taxes last year?	
If all of our auditing clients want their auditing work done by 12/31, how will we organize our audit work in the new year?	
Can the IRS use predictive techniques to find those individuals or corporations that are evading taxes?	
Did the airline company's on-time departures improve this past month?	
Can our variance analysis help explain why our labor expenses increased over the past year?	

5. (LO 8.1, 8.4, 8.5, 8.6, 8.7) Match the analytics type (descriptive, diagnostic, predictive, or prescriptive) and branch of accounting (financial, managerial, audit, or tax) to the analytics technique used.

Analytics Technique Used	Analytics Type (Descriptive, Diagnostic, Predictive, or Prescriptive)	Branch of Accounting (Financial, Managerial, Audit, or Tax)
Classification to accept or reject audit clients		
Time series analysis to forecast sales		
Common-size financial statements		
Regression to assess the company's cost structure (determine fixed costs and variable costs)		
Goal-seek analysis to identify the break-even point		
Sequence checks assessing missing transactions		
Vertical analysis for investors to compare a company to its peers		

6. (LO 8.4, 8.5, 8.6, 8.7) Match each type of analysis to its analytics type (descriptive, diagnostic, predictive, or prescriptive).

Types of Analysis	**Analytics Type** (Descriptive, Diagnostic, Predictive, or Prescriptive)
Ratio analytics	
Bank reconciliation	
Scenario analytics	
Pivot table analytics to find the most profitable products or customers	
Fuzzy matching of vendors and employees to see if there are similarities in names or addresses	
Cash flow analysis	
Benford's Law	

7. (LO 8.4) Match the business analytics technique included in Excel's Data Analysis ToolPak to its use.

Business Analytics Technique Included in Excel's Data Analysis ToolPak	**Use**
Correlation	To understand the frequency of the data using a display of rectangles with areas proportional to the underlying frequency of the data
Regression	To understand if and the extent to which variables are related to each other
Descriptive statistics	To query, organize, and summarize data to make comparisons and see patterns or trends in the data
Histogram	To understand the basic statistics, including the mean, standard deviation, minimum, and maximum of a data set
Pivot tables	To understand the relation between specific outputs and variable inputs

Problems

1. (LO 8.1, 8.4, 8.5, 8.6, 8.7) Describe specific types of questions asked in *auditing* that might require descriptive, diagnostic, predictive, and prescriptive analytics (one question for each). Also describe the types of specific tools and techniques that could be used to address each of those questions.
2. (LO 8.1, 8.4, 8.5, 8.6, 8.7) Describe specific types of questions asked in *tax accounting* that might require descriptive, diagnostic, predictive, and prescriptive analytics (one question for each). Also describe the types of specific tools and techniques that could be used to address each of those questions.
3. (LO 8.5) You're an auditor and you're looking for fraudulent transactions.

Required

a. If you had access to any accounting records you wanted (with details of every transaction), where would you look? Explain.

b. How would you find kickbacks to employees from vendors? Where would you look? (See Exhibit 8.20 for an example.)

c. What type of diagnostic analytics would you use?

4. (LO 8.4) Refer to Exhibit 8.19. Is **J.B. Hunt**'s performance improving or declining overall? Why is it improving or declining in the most recent year? Which specific line items help you answer these questions?
5. (LO 8.4) DJI Mavic is attempting to understand its sales and cost structure to allocate appropriate personnel to its sales force. Download the DJI Mavic Drone Sales Journal data set in Excel. Note the sales price and cost of each sale. Either use the descriptive statistics tool featured in Excel's Analysis ToolPak or Excel functions to compute the summary statistics and answer the following questions.

Required

a. What is the average sales price for all items sold?
b. What is the standard deviation of the sales price for all items sold?
c. What is the range of costs for all items sold?
d. What is the lowest cost for all items sold?

6. (LO 8.4) SkyDio is attempting to understand its sales and cost structure to understand its overall profitability in relation to each sale. Download the SkyDio Drone Sales Journal data set in Excel. Note the sales price and cost of each sale. In Excel, calculate the gross margin percentage (Sales Price – Cost)/Sales Price for each sale (or row in the sales journal). Use either the descriptive statistics tool featured in Excel's Data Analysis ToolPak or Excel functions to compute the summary statistics and answer the following questions.

Required

a. What is the average gross margin percentage for all items sold?
b. What is the standard error of the gross margin percentage for all items sold?
c. What is the range of gross margin percentage for all items sold?
d. What is the lowest gross margin percentage for all items sold?
e. What is the count of sales transactions in the data set?

7. (LO 8.7) Management accountants can use the following table to perform a sensitivity analysis, evaluating the company's changing profitability based on two important inputs: (1) the level of total fixed costs and (2) the sales price of the product. Note that the analysis varies the sales price from $35 to $53 (as shown in the columns) and the fixed costs from $500 to $3,000 (as shown in the rows). Total sales are constant at 300 units and the variable costs are constant at $20 per unit.

   The colored columns are the total profits based on the changing input parameters.

		Sales Price									
		$ 35.00	$ 37.00	$ 39.00	$ 41.00	$ 43.00	$ 45.00	$ 47.00	$ 49.00	$ 51.00	$ 53.00
Fixed Costs	500	4000.00	4600.00	5200.00	5800.00	6400.00	7000.00	7600.00	8200.00	8800.00	9400.00
	750	3750.00	4350.00	4950.00	5550.00	6150.00	6750.00	7350.00	7950.00	8550.00	9150.00
	1,000	3500.00	4100.00	4700.00	5300.00	5900.00	6500.00	7100.00	7700.00	8300.00	8900.00
	1,250	3250.00	3850.00	4450.00	5050.00	5650.00	6250.00	6850.00	7450.00	8050.00	8650.00
	1,500	3000.00	3600.00	4200.00	4800.00	5400.00	6000.00	6600.00	7200.00	7800.00	8400.00
	1,750	2750.00	3350.00	3950.00	4550.00	5150.00	5750.00	6350.00	6950.00	7550.00	8150.00
	2,000	2500.00	3100.00	3700.00	4300.00	4900.00	5500.00	6100.00	6700.00	7300.00	7900.00
	2,250	2250.00	2850.00	3450.00	4050.00	4650.00	5250.00	5850.00	6450.00	7050.00	7650.00
	2,500	2000.00	2600.00	3200.00	3800.00	4400.00	5000.00	5600.00	6200.00	6800.00	7400.00
	2,750	1750.00	2350.00	2950.00	3550.00	4150.00	4750.00	5350.00	5950.00	6550.00	7150.00
	3,000	1500.00	2100.00	2700.00	3300.00	3900.00	4500.00	5100.00	5700.00	6300.00	6900.00

Required

a. As sales price increases, what happens to the estimated profitability?
b. As total fixed costs increase, what happens to the estimated profitability?
c. If the sales price is $53 and total fixed costs are $3,000, what is the estimated profitability?
d.

8. (LO 8.7) Using the exhibit and description in Problem 7 above, evaluate the sensitivity of profits to changing levels of fixed costs and product sales price.

Required

a. Which has a greater impact on profitability, a $2 increase in sales price or a $250 decrease in total fixed costs?

b. How does conditional formatting help you to visualize the impact of these two important inputs on estimated profitability?

c. Would you recommend a different coloring scheme than the red-yellow-green scheme? Why or why not?

d. What other inputs do you think would be relevant to evaluate in assessing profitability?

9. (LO 8.5) Download the Costs of Production data set in Excel. Note the Units of Production (labeled Production) in Column B and the Total Costs (labeled Cost) by week.

Required

a. Perform a regression analysis with Cost as the dependent (outcome) variable and Production as the independent variable.

b. Estimate the adjusted R-square and Model F statistic of the regression output to assess the model fit.

c. Using the regression parameters, estimate the level of fixed costs overall and the variable costs overall.



DATA VISUALIZATION

Data Visualization Exercise

1. (LO 8.4) Download the SkyDio Drone Sales Journal data set in Excel. Note the sales price and cost of each sale. In Excel, calculate the gross margin percentage (Sales Price – Cost)/Sales Price for each sale (or row in the sales journal). Use either the descriptive statistics tool featured in Excel's Data Analysis ToolPak or Excel functions to compute the summary statistics and answer the following questions.

Required

a. Create a histogram of the gross margin percentages. (*Hint:* **Insert** > **Charts** > **Statistic Chart** > **Histogram** after selecting the data range.) Submit this screenshot to your instructor, if assigned.

b. How many transactions were for 47% or higher?

c. How many transactions were for 40% or lower?

LAB 8.1 **Excel:** Descriptive Analytics: Performing Horizontal Analysis
LAB 8.2 **Excel:** Descriptive Analytics: Performing Vertical Analysis
LAB 8.3 **Excel:** Diagnostic Analytics: Using Fuzzy Matching to Look for Fraud
LAB 8.4 **Excel:** Diagnostic Analytics: Estimating Fixed and Variable Costs
LAB 8.5 **Excel:** Predictive Analytics: Forecasting Future Performance of IBM and Netflix
LAB 8.6 **Tableau:** Predictive Analytics: Forecasting Future Performance of IBM and Netflix
LAB 8.7 **Power BI:** Predictive Analytics: Forecasting Future Performance of IBM and Netflix
LAB 8.8 **Excel:** Prescriptive Analytics: Using Goal Seek Analysis to Determine the Break-Even Point

Lab 8.1 Excel

Descriptive Analytics: Performing Horizontal Analysis

tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keyword

Financial Performance

Decision-Making Context

Several types of financial statement analysis can be performed as part of accounting analytics, including *horizontal analysis,* which calculates and displays the dollar and percentage change in financial statements from one period to the next. Horizontal analysis provides comparative information about various line items of the income statement, balance sheet, or statement of cash flows.

Required

Perform horizontal analysis on **Apple**'s income statement.

Specify the Question

How can horizontal analysis of Apple's income statement highlight changes from one year to the next?

Obtain the Data

1. Open Excel file Lab 8.1 Data.xlsx.
2. Browse the spreadsheet that includes Apple's income statement over 2020 and 2021. Lab Exhibit 8.1.1 is a snippet of the spreadsheet.

Lab Exhibit 8.1.1

	A	B	C
1	Apple Inc. (AAPL)		
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)		
3	*Fiscal Period*	**Y 2020**	**Y 2021**
4	*Period Start*	**9/29/2019**	**9/27/2020**
5	*Period End*	**9/26/2020**	**9/25/2021**
6	Products	$220,747	$297,392
7	Services	$53,768	$68,425
8	Net sales	$274,515	$365,817
9	Products	$151,286	$192,266
10	Services	$18,273	$20,715
11	Cost of sales	$169,559	$212,981
12	Gross margin	$104,956	$152,836
13	**Operating expenses:**		
14	Research and development	$18,752	$21,914
15	Selling, general and administrative	$19,916	$21,973
16	Total operating expenses	$38,668	$43,887
17	Operating income	$66,288	$108,949
18	Other income/(expense), net	$803	$258
19	Income before provision for income taxes	$67,091	$109,207
20	Provision for income taxes	$9,680	$14,527
21	Net income	$57,411	$94,680

Analyze the Data

Analysis Task 1: Prepare the Spreadsheet.

1. In cell D4, input the words "Change from" in cell D5 and insert the words "2020 to 2021", as shown in Lab Exhibit 8.1.2. Boldface both cells D4 and D5.

Lab Exhibit 8.1.2

	A	B	C	D
1	Apple Inc. (aapl)			
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)			
3	*Fiscal Period*	**Y 2020**	**Y 2021**	
4	*Period Start*	**9/29/2019**	**9/27/2020**	**Change from**
5	*Period End*	**9/26/2020**	**9/25/2021**	**2020 to 2021**
6	Products	$220,747	$297,392	
7	Services	$53,768	$68,425	

2. Starting in cell D6, we'll compute the dollar amount of change from 2020 to 2021. Take the newer year (2021) and subtract the older year (2020), as shown in Lab Exhibit 8.1.3.

Lab Exhibit 8.1.3

	A	B	C	D
1	Apple Inc. (aapl)			
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)			
3	*Fiscal Period*	**Y 2020**	**Y 2021**	
4	*Period Start*	**9/29/2019**	**9/27/2020**	**Change from**
5	*Period End*	**9/26/2020**	**9/25/2021**	**2020 to 2021**
6	Products	$220,747	$297,392	=C6-B6

3. Copy the result of cell D6 to all the relevant cells in Column D to get the result shown in Lab Exhibit 8.1.4.

Lab Exhibit 8.1.4

	A	B	C	D
1	Apple Inc. (aapl)			
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)			
3	*Fiscal Period*	**Y 2020**	**Y 2021**	
4	*Period Start*	**9/29/2019**	**9/27/2020**	**Change from**
5	*Period End*	**9/26/2020**	**9/25/2021**	**2020 to 2021**
6	Products	$220,747	$297,392	$76,645
7	Services	$53,768	$68,425	$14,657
8	Net sales	$274,515	$365,817	$91,302
9	Products	$151,286	$192,266	$40,980
10	Services	$18,273	$20,715	$2,442
11	Cost of sales	$169,559	$212,981	$43,422
12	Gross margin	$104,956	$152,836	$47,880
13	**Operating expenses:**			
14	Research and development	$18,752	$21,914	$3,162
15	Selling, general and administrative	$19,916	$21,973	$2,057
16	Total operating expenses	$38,668	$43,887	$5,219
17	Operating income	$66,288	$108,949	$42,661
18	Other income/(expense), net	$803	$258	($545)
19	Income before provision for income taxes	$67,091	$109,207	$42,116
20	Provision for income taxes	$9,680	$14,527	$4,847
21	Net income	$57,411	$94,680	$37,269

4. Copy the formatting of Column C to Column D. To do so, highlight Column C, select the Format Painter button, and then highlight Column D to complete the format copying.

Analysis Task 2: Compute the Percentage Change.

1. First, let's enter titles. In cell E3, type the word "Percentage". In cell E4, type the words "Change from". In cell E4, type the words "2020 to 2021".
2. Next, let's compute the percentage change by taking the most current (2021), subtract the less current (2020), and divide by the less current, as in the following formula:

$$\text{Percentage change (increase or decrease)} = \frac{(\text{New} - \text{Old}) * 100}{\text{Old}}$$

To calculate the percentage, in cell E6 enter "=D6/B6" (Lab Exhibit 8.1.5).

Lab Exhibit 8.1.5

	A	B	C	D	E
1	Apple Inc. (aapl)				
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)				
3	*Fiscal Period*	**Y 2020**	**Y 2021**		**Percentage**
4	*Period Start*	**9/29/2019**	**9/27/2020**	**Change from**	**Change from**
5	*Period End*	**9/26/2020**	**9/25/2021**	**2020 to 2021**	**2020 to 2021**
6	Products	$220,747	$297,392	$76,645	=D6/B6

3. Next, we need to format the cells in Column E as percentages. Highlight Column E and then right-click on **Format Cells > Percentage** (for one decimal place). Then click OK.
4. Copy the result of cell E6 to all relevant cells below it in Column E. The spreadsheet should now appear as shown in Lab Exhibit 8.1.6.

Lab Exhibit 8.1.6

	A	B	C	D	E
1	Apple Inc. (aapl)				
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)				
3	*Fiscal Period*	Y 2020	Y 2021		Percentage
4	*Period Start*	9/29/2019	9/27/2020	Change from	Change from
5	*Period End*	9/26/2020	9/25/2021	2020 to 2021	2020 to 2021
6	Products	$220,747	$297,392	$76,645	34.7%
7	Services	$53,768	$68,425	$14,657	27.3%
8	Net sales	$274,515	$365,817	$91,302	33.3%
9	Products	$151,286	$192,266	$40,980	27.1%
10	Services	$18,273	$20,715	$2,442	13.4%
11	Cost of sales	$169,559	$212,981	$43,422	25.6%
12	Gross margin	$104,956	$152,836	$47,880	45.6%
13	**Operating expenses:**				
14	Research and development	$18,752	$21,914	$3,162	16.9%
15	Selling, general and administrative	$19,916	$21,973	$2,057	10.3%
16	Total operating expenses	$38,668	$43,887	$5,219	13.5%
17	Operating income	$66,288	$108,949	$42,661	64.4%
18	Other income/(expense), net	$803	$258	($545)	-67.9%
19	Income before provision for income taxes	$67,091	$109,207	$42,116	62.8%
20	Provision for income taxes	$9,680	$14,527	$4,847	50.1%
21	Net income	$57,411	$94,680	$37,269	64.9%

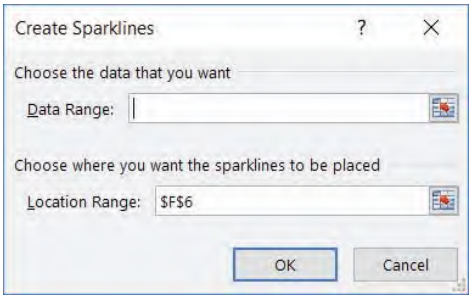


DATA VISUALIZATION

Analysis Task 3: Add Sparkline Visualizations.

1. The final task is to add sparklines to get an idea of the relative increase or decrease in a graphic form. To do so, go to cell F6, and then in the ribbon, click on **Insert > Sparklines > Line**. The dialog box in Lab Exhibit 8.1.7 will appear.

Lab Exhibit 8.1.7



2. Within the dialog box, insert "B6:C6" (or highlight cells B6 and C6). Copy cell F6 to all relevant cells in Column F.
3. The resulting spreadsheet should appear as in Lab Exhibit 8.1.8 and is ready to share with decision-makers.

Report the Results

We have now completed the horizontal analysis, and we can share it with decision-makers to show them which line items increased, which line items decreased, and by

Lab Exhibit 8.1.8

	A	B	C	D	E	F
1	Apple Inc. (aapl)					
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)					
3	*Fiscal Period*	Y 2020	Y 2021		Percentage	
4	*Period Start*	9/29/2019	9/27/2020	Change from	Change from	
5	*Period End*	9/26/2020	9/25/2021	2020 to 2021	2020 to 2021	
6	Products	$220,747	$297,392	$76,645	34.7%	
7	Services	$53,768	$68,425	$14,657	27.3%	
8	Net sales	$274,515	$365,817	$91,302	33.3%	
9	Products	$151,286	$192,266	$40,980	27.1%	
10	Services	$18,273	$20,715	$2,442	13.4%	
11	Cost of sales	$169,559	$212,981	$43,422	25.6%	
12	Gross margin	$104,956	$152,836	$47,880	45.6%	
13	**Operating expenses:**					
14	Research and development	$18,752	$21,914	$3,162	16.9%	
15	Selling, general and administrative	$19,916	$21,973	$2,057	10.3%	
16	Total operating expenses	$38,668	$43,887	$5,219	13.5%	
17	Operating income	$66,288	$108,949	$42,661	64.4%	
18	Other income/(expense), net	$803	$258	($545)	-67.9%	
19	Income before provision for income taxes	$67,091	$109,207	$42,116	62.8%	
20	Provision for income taxes	$9,680	$14,527	$4,847	50.1%	
21	Net income	$57,411	$94,680	$37,269	64.9%	

what percentage. This information will provide them with more knowledge of what happened over the past two years. Even though revenues increased by just 33.3%, net income increased by 64.9%.

Assessment

1. Take a screenshot of the resulting horizontal analysis, paste it into a Word document named "Lab 8.1 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Lab 8.1 Excel Submission.docx via Connect, if assigned.

Alternate Lab 8.1 Excel

Descriptive Analytics: Performing Horizontal Analysis

Apply the same steps in Lab 8.1 to the Alt Lab 8.1 Data.xlsx data set to perform horizontal analysis for **Google**.

Required

Perform horizontal analysis on Google's income statement.

Open Excel file Alt Lab 8.1 Data.xlsx.

Assessment

1. Take a screenshot of the resulting horizontal analysis, paste it into a Word document named "Alt Lab 8.1 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 8.1 Excel Submission.docx via Connect, if assigned.

Lab 8.2 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Descriptive Analytics: Performing Vertical Analysis

Keyword

Financial Performance

Decision-Making Context

Several types of financial statement analysis can be performed to better understand a company's financial position and financial performance. *Vertical analysis* expresses financial information in relation to a base amount. For income-statement items, the relevant base is net sales revenue. On balance-sheet items, the relevant base is total assets.

Required

Perform vertical analysis on the **Apple Inc**. income statement.

Specify the Question

How can vertical analysis of the Apple Inc. income statement highlight changes from one year to the next?

Obtain the Data

1. Open Excel file Lab 8.2 Data.xlsx.
2. Browse the spreadsheet that includes Apple Inc.'s income statement over 2020 and 2021. Lab Exhibit 8.2.1 provides a snippet of the spreadsheet.

Lab Exhibit 8.2.1

	A	B	C
1	Apple Inc. (aapl)		
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)		
3	*Fiscal Period*	**Y 2020**	**Y 2021**
4	*Period Start*	**9/29/2019**	**9/27/2020**
5	*Period End*	**9/26/2020**	**9/25/2021**
6	Products	$220,747	$297,392
7	Services	$53,768	$68,425
8	Net sales	$274,515	$365,817
9	Products	$151,286	$192,266
10	Services	$18,273	$20,715
11	Cost of sales	$169,559	$212,981
12	Gross margin	$104,956	$152,836
13	**Operating expenses:**		
14	Research and development	$18,752	$21,914
15	Selling, general and administrative	$19,916	$21,973
16	Total operating expenses	$38,668	$43,887
17	Operating income	$66,288	$108,949
18	Other income/(expense), net	$803	$258
19	Income before provision for income taxes	$67,091	$109,207
20	Provision for income taxes	$9,680	$14,527
21	Net income	$57,411	$94,680

Analyze the Data

Analysis Task 1: Compute Each Line Item as a Percentage of Sales.

1. In cell D4, input the words "Percent of Sales" (Lab Exhibit 8.2.2). Highlight cells D4 and E4, and click on the ribbon, **Home > Alignment > Merge & Center**.

Lab Exhibit 8.2.2

	A	B	C	D	E
1	Apple Inc. (aapl)				
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)				
3	*Fiscal Period*	Y 2020	Y 2021		
4	*Period Start*	9/29/2019	9/27/2020	Percent of Sales	

2. In cell D5, type "2020" and in cell E5, type "2021".
3. We will scale/deflate (or divide) all variables by the net sales of each respective year on each line.
   For the 2020 income statement figures, we will begin our vertical analysis. In cell D6, insert the formula "=B6/$B$8". We use the absolute reference for 2020 net sales so that we can copy the formula to other cells and still have it refer to 2020 net sales (Lab Exhibit 8.2.3).

Lab Exhibit 8.2.3

	A	B	C	D
1	Apple Inc. (aapl)			
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)			
3	*Fiscal Period*	Y 2020	Y 2021	
4	*Period Start*	9/29/2019	9/27/2020	Percent c
5	*Period End*	9/26/2020	9/25/2021	2020
6	Products	$220,747	$297,392	=B6/$B$8
7	Services	$53,768	$68,425	
8	Net sales	$274,515	$365,817	

4. Copy the contents of cell D6 to all relevant cells in Column D.
5. For the 2021 income statement figures, we will begin our vertical analysis. In cell E6, insert the formula "=C6/$C$8". We use the absolute reference for 2021 net sales so that we can copy the formula to other cells and still have it refer to 2021 net sales.
6. Copy the contents of cell E6 to all relevant cells in Column E.
7. Next, we need to format the cells in Columns D and E as percentages. Highlight cells D6:E21 and then right-click on **Format Cells > Percentage** (for two decimal places). Then click OK to get the resulting spreadsheet, as shown in Lab Exhibit 8.2.4.

Lab Exhibit 8.2.4

	A	B	C	D	E
1	Apple Inc. (aapl)				
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)				
3	*Fiscal Period*	Y 2020	Y 2021		
4	*Period Start*	9/29/2019	9/27/2020	Percent of Sales	
5	*Period End*	9/26/2020	9/25/2021	2020	2021
6	Products	$220,747	$297,392	80.41%	81.30%
7	Services	$53,768	$68,425	19.59%	18.70%
8	Net sales	$274,515	$365,817	100.00%	100.00%
9	Products	$151,286	$192,266	55.11%	52.56%
10	Services	$18,273	$20,715	6.66%	5.66%
11	Cost of sales	$169,559	$212,981	61.77%	58.22%
12	Gross margin	$104,956	$152,836	38.23%	41.78%
13	**Operating expenses:**				
14	Research and development	$18,752	$21,914	6.83%	5.99%
15	Selling, general and administrative	$19,916	$21,973	7.25%	6.01%
16	Total operating expenses	$38,668	$43,887	14.09%	12.00%
17	Operating income	$66,288	$108,949	24.15%	29.78%
18	Other income/(expense), net	$803	$258	0.29%	0.07%
19	Income before provision for income taxes	$67,091	$109,207	24.44%	29.85%
20	Provision for income taxes	$9,680	$14,527	3.53%	3.97%
21	Net income	$57,411	$94,680	20.91%	25.88%

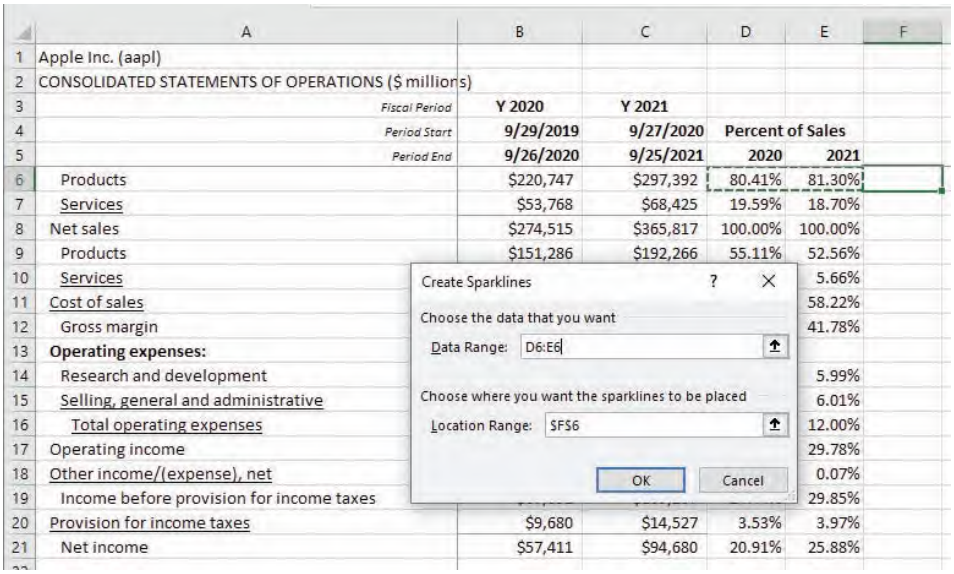


DATA VISUALIZATION

Analysis Task 2: Add Sparkline Visualizations.

1. The final piece of our analysis requires us to add sparklines to get an idea of the relative increase or decrease in a graphic form. To do so, go to cell F6, and then in the ribbon, select **Insert > Sparklines > Line**. The dialog box shown in Lab Exhibit 8.2.5 will appear.

Lab Exhibit 8.2.5



	A	B	C	D	E	F
1	Apple Inc. (aapl)					
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)					
3	*Fiscal Period*	Y 2020	Y 2021			
4	*Period Start*	9/29/2019	9/27/2020	Percent of Sales		
5	*Period End*	9/26/2020	9/25/2021	2020	2021	
6	Products	$220,747	$297,392	80.41%	81.30%	
7	Services	$53,768	$68,425	19.59%	18.70%	
8	Net sales	$274,515	$365,817	100.00%	100.00%	
9	Products	$151,286	$192,266	55.11%	52.56%	
10	Services				5.66%	
11	Cost of sales				58.22%	
12	Gross margin				41.78%	
13	**Operating expenses:**					
14	Research and development				5.99%	
15	Selling, general and administrative				6.01%	
16	Total operating expenses				12.00%	
17	Operating income				29.78%	
18	Other income/(expense), net				0.07%	
19	Income before provision for income taxes				29.85%	
20	Provision for income taxes	$9,680	$14,527	3.53%	3.97%	
21	Net income	$57,411	$94,680	20.91%	25.88%	

2. Within the dialog box, insert "D6:E6" (or highlight cells D6 and E6). Copy cell F6 to all relevant cells in Column F. The resulting spreadsheet should appear as shown in Lab Exhibit 8.2.6 and is ready to share with decision-makers.

Lab Exhibit 8.2.6

	A	B	C	D	E	F
1	Apple Inc. (aapl)					
2	CONSOLIDATED STATEMENTS OF OPERATIONS ($ millions)					
3	*Fiscal Period*	Y 2020	Y 2021			
4	*Period Start*	9/29/2019	9/27/2020	Percent of Sales		
5	*Period End*	9/26/2020	9/25/2021	2020	2021	
6	Products	$220,747	$297,392	80.41%	81.30%	
7	Services	$53,768	$68,425	19.59%	18.70%	
8	Net sales	$274,515	$365,817	100.00%	100.00%	
9	Products	$151,286	$192,266	55.11%	52.56%	
10	Services	$18,273	$20,715	6.66%	5.66%	
11	Cost of sales	$169,559	$212,981	61.77%	58.22%	
12	Gross margin	$104,956	$152,836	38.23%	41.78%	
13	**Operating expenses:**					
14	Research and development	$18,752	$21,914	6.83%	5.99%	
15	Selling, general and administrative	$19,916	$21,973	7.25%	6.01%	
16	Total operating expenses	$38,668	$43,887	14.09%	12.00%	
17	Operating income	$66,288	$108,949	24.15%	29.78%	
18	Other income/(expense), net	$803	$258	0.29%	0.07%	
19	Income before provision for income taxes	$67,091	$109,207	24.44%	29.85%	
20	Provision for income taxes	$9,680	$14,527	3.53%	3.97%	
21	Net income	$57,411	$94,680	20.91%	25.88%	

Report the Results

Now that we have completed the vertical analysis, we can share it with decision-makers to show them which items increased, which items decreased, and by what percentages, helping them better understand what happened.

Assessment

1. Take a screenshot of the resulting vertical analysis, paste it into a Word document named "Lab 8.2 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Lab 8.2 Excel Submission.docx via Connect, if assigned.

Alternate Lab 8.2 Excel

Descriptive Analytics: Performing Vertical Analysis

Apply the same steps in Lab 8.2 to the Alt Lab 8.2 Data.xlsx data set to perform vertical analysis for **Google**.

Required

1. Perform vertical analysis on Google's income statement.
2. Open Excel file Alt Lab 8.2 Data.xlsx.

Assessment

1. Take a screenshot of the resulting vertical analysis, paste it into a Word document named "Alt Lab 8.2 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 8.2 Excel Submission.docx via Connect, if assigned.

Lab 8.3 Excel

Lab Note 1: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Lab Note 2: You will need the Fuzzy Lookup add-in to complete this lab. There are details on how to find and install the add-in below.

Diagnostic Analytics: Using Fuzzy Matching to Look for Fraud

Keywords

Fuzzy Match, Data Reduction, Fraud Detection

Decision-Making Context

Auditors and data analysts use data-reduction techniques to narrow down transactions from a big data set of transactions to a smaller data set that might contain fraud or errors. This technique helps them effectively decrease their effort and increase their work efficiency by allowing them to examine the transactions that might require more attention.

A widely used approach in data reduction is *fuzzy matching,* which allows analysts to merge two data sets using the similarity of a common variable. Users can set a specific threshold for the similarity of the common variable of each data set to merge both data sets. In addition, they can apply fuzzy matching to combine data sets generated by different accounting information systems. For example, my name in one data set might be Vern Richardson, and in another data set, Vernon Jay Richardson. The use of a fuzzy match would allow my name to be the common variable, and information from both data sets might be merged into one data set.

In this lab, we take the perspective of an auditor who wants to reduce data by (1) highlighting suspicious transactions that are possibly made by company employees and (2) merging data sets from different accounting information systems.

Required

Use fuzzy matching to detect suspicious transactions.

Specify the Question

How can fuzzy matching be used to identify suspicious transactions?

Obtain the Data

Premier Insurance Ltd. is a (fictional) insurance company. Because December earnings barely met the earnings target expected by shareholders and financial analysts, the internal auditors think that fraud may be occurring.

1. Open Excel file Lab 8.3 Data.xlsx. The data dictionary appears below.
2. Browse the spreadsheet to make sure there aren't any obvious errors in the Excel file.

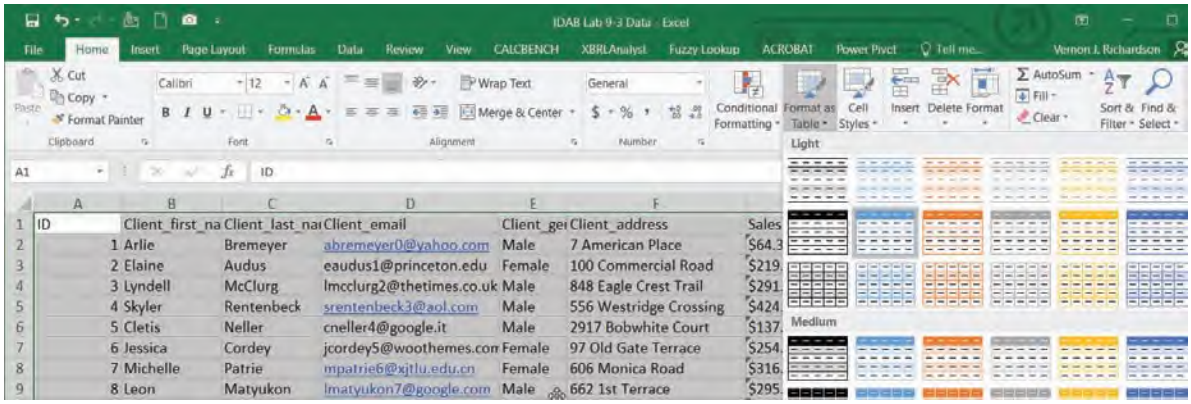
Data Dictionary

ID:	The identify key of each transaction
Client_first_name:	The first name of the client/employee
Client_last_name:	The last name of the client/employee
Client_Email:	The email address of the client/employee
Gender:	The gender of the client/employee
Address:	The address of the client/employee
Sales:	Dollar amount of sales revenue from each transaction

Analyze the Data

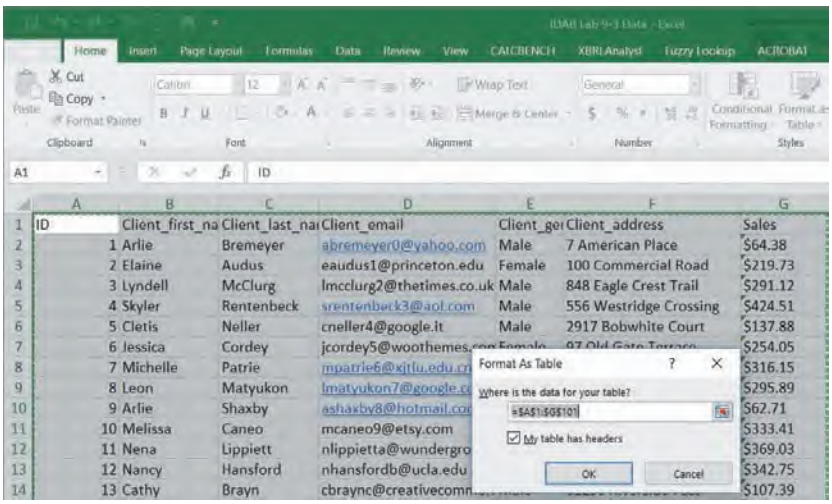
Analysis Task 1: Prepare the Data for Fuzzy Matching.

1. Go to the spreadsheet tab Client_Info. Select all the records. Click the top-left cell, press Shift, and then click the bottom-right cell.
2. Go to the Home tab in the menu bar. In the Styles section, click Format as Table and select a style that you feel presents data most clearly (Lab Exhibit 8.3.1).



Lab Exhibit 8.3.1

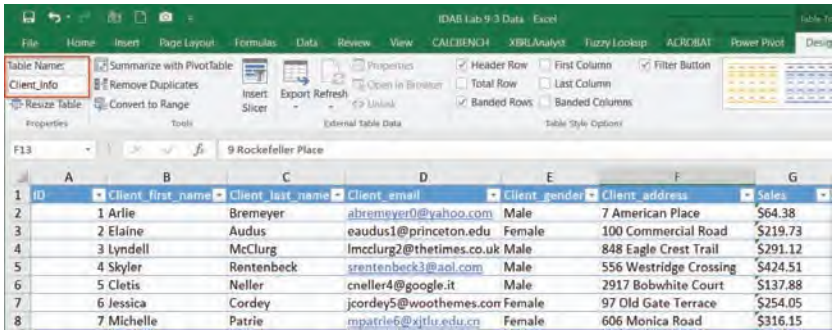
3. In the window of Format As Table, check the box next to My table has headers, and click OK (Lab Exhibit 8.3.2).



Lab Exhibit 8.3.2

4. Rename the table. Go to Table Tools – Design. In the Properties section, under Table name, rename table 1 as “Client_Info” (Lab Exhibit 8.3.3).

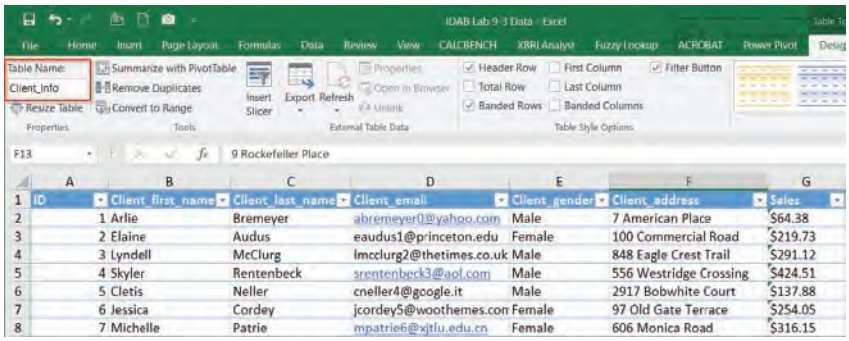
Lab Exhibit 8.3.3



ID	Client_first_name	Client_last_name	Client_email	Client_gender	Client_address	Sales
1	Arlie	Bremeyer	abremeyer0@yahoo.com	Male	7 American Place	$64.38
2	Elaine	Audus	eaudus1@princeton.edu	Female	100 Commercial Road	$219.73
3	Lyndell	McClurg	lmcclurg2@thetimes.co.uk	Male	848 Eagle Crest Trail	$291.12
4	Skyler	Rentenbeck	srentenbeck3@aol.com	Male	556 Westridge Crossing	$424.51
5	Cletis	Neller	cneller4@google.it	Male	2917 Bobwhite Court	$137.88
6	Jessica	Cordey	jcordey5@woothemes.com	Female	97 Old Gate Terrace	$254.05
7	Michelle	Patrie	mpatrie6@xjtlu.edu.cn	Female	606 Monica Road	$316.15

5. Go to the spreadsheet Employee_Info and follow the same Steps 3 to 6 above. Then name the table "Employee_Info" (Lab Exhibit 8.3.4).

Lab Exhibit 8.3.4

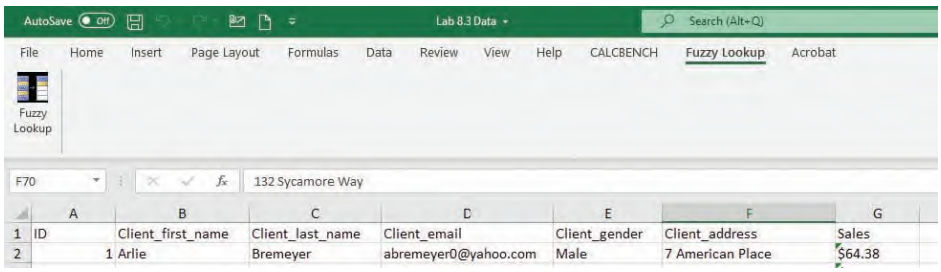


ID	Client_first_name	Client_last_name	Client_email	Client_gender	Client_address	Sales
1	Arlie	Bremeyer	abremeyer0@yahoo.com	Male	7 American Place	$64.38
2	Elaine	Audus	eaudus1@princeton.edu	Female	100 Commercial Road	$219.73
3	Lyndell	McClurg	lmcclurg2@thetimes.co.uk	Male	848 Eagle Crest Trail	$291.12
4	Skyler	Rentenbeck	srentenbeck3@aol.com	Male	556 Westridge Crossing	$424.51
5	Cletis	Neller	cneller4@google.it	Male	2917 Bobwhite Court	$137.88
6	Jessica	Cordey	jcordey5@woothemes.com	Female	97 Old Gate Terrace	$254.05
7	Michelle	Patrie	mpatrie6@xjtlu.edu.cn	Female	606 Monica Road	$316.15

Analysis Task 2: Perform Fuzzy Matching.

1. If you have never used Fuzzy Lookup add-in Excel, go to https://www.microsoft.com/en-us/download/details.aspx?id=15011&msclkid=3dfb37e2abc411ecad6c4c42dbc3daa2 to download and install the Fuzzy Lookup add-in for Excel. This add-in is available only for Windows-based computers.
2. Once you've installed the add-in, go to Fuzzy Lookup in the ribbon and click the Fuzzy Lookup icon. A panel with a header of Fuzzy Lookup will show up on the right (Lab Exhibit 8.3.5).

Lab Exhibit 8.3.5

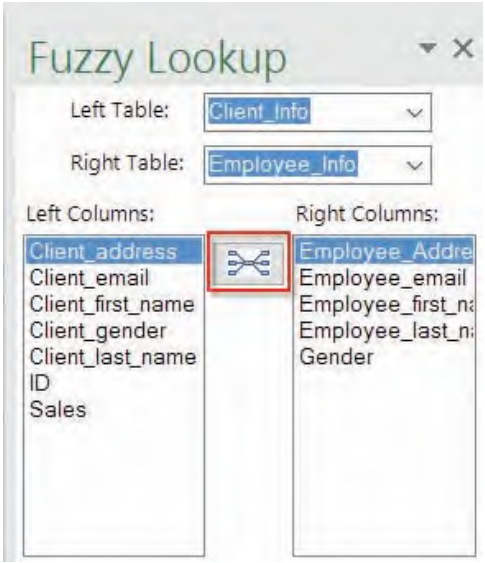


ID	Client_first_name	Client_last_name	Client_email	Client_gender	Client_address	Sales
1	Arlie	Bremeyer	abremeyer0@yahoo.com	Male	7 American Place	$64.38

3. For the Left Table, select Client_Info, and for the Right Table, choose Employee_Info.

4. Under Left Columns, select Client_address. Under Right Columns, select Employee_address.
5. Click the merge icon in the middle (Lab Exhibit 8.3.6).

Lab Exhibit 8.3.6



6. Under Output Columns, check boxes next to:
   - Client_first_name
   - Client_last_name
   - Client_address
   - Employee_first_name
   - Employee_last_name
   - Employee_address
   - FuzzyLookup.Similarity
7. For Number of Matches, choose 1. Similarity threshold represents the percentage similarity in the range from 0 to 1. Zero means "completely different" and one means "exactly the same." You can slide the slider to the left or right to see the difference. Set the similarity threshold to 0.50 (Lab Exhibit 8.3.7).
8. Go to a new worksheet, click cell A1, then click Go in the panel of Fuzzy Lookup.
9. Name the worksheet "FuzzyMatch".
10. Format the worksheet as a Table and use Filter to exclude data records with 0 similarity. Select all the records, copy and paste records in a new worksheet tab, and name it FuzzyMatchClean.

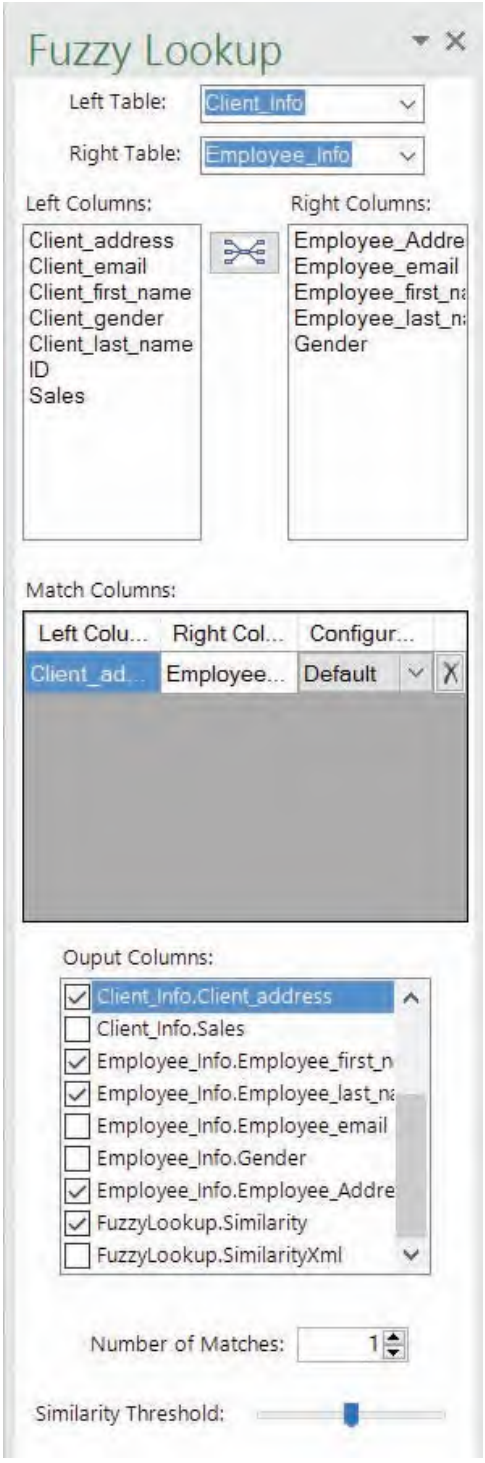
Report the Results

After we extract suspicious transactions, we can do more research on these transactions to see if they might be fraudulent. For example, we might use an internet search to see if there is an actual business at that address and talk to employees to learn more details about the transactions.

Assessment

1. Take a screenshot of the Fuzzy Lookup Control Panel (where you set the specifications for fuzzy lookup), paste it into a Word document named "Lab 8.3 Excel Submission.docx", and label the screenshot Submission 1.

Lab Exhibit 8.3.7



2. Take a screenshot of the worksheet tab FuzzyMatchClean, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 8.3 Excel Submission.docx via Connect, if assigned.

Alternate Lab 8.3 Excel

Diagnostic Analytics: Using Fuzzy Matching to Look for Fraud

Apply the same steps in Lab 8.3 to the Alt Lab 8.3 Data.xlsx data set. Instead of asking you to detect fraudulent transactions, this lab asks you to merge two data sets collected from two different information systems. Specifically, you will compare the client system with the employee system to see if there are any close matches.

Join two worksheets using Fuzzy Matching in Alt Lab 8.3 Data.xlsx by *Company name*. Set the similarity threshold to 0.50 and output all of system1 and system2 fields.

Required

Use fuzzy matching to detect suspicious transactions.

Assessment

1. Take a screenshot of the Fuzzy Matching Control Panel (where you set the specifications for fuzzy lookup), paste it into a Word document named "Alt Lab 8.3 Excel Submission.docx", and label the screenshot Submission 1.
2. Filter out the records with a similarity lower than 0.90, take a screenshot of the filtered table, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 8.3 Excel Submission.docx via Connect, if assigned.

Lab 8.4 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Diagnostic Analytics: Estimating Fixed and Variable Costs

Keywords

Managerial Accounting, Cost Behavior, Fixed Costs, Variable Costs

Decision-Making Context

To successfully run a business, we need to understand our costs. *Fixed costs* do not change with the number of goods or services produced. *Variable costs* vary with the level of output.

Many companies conduct various types of analyses, including regression analyses, to get a better handle on their cost behavior structure (that is, the proportion of fixed costs and the proportion of variable costs). We will use regression analysis in this lab.

Required

1. Use regression analysis to estimate fixed and variable costs.
2. Use regression analysis to determine if costs are different during the busy season.
3. Use regression analysis to determine if the company is better off incorporating the busy season into its cost behavior analysis.

Specify the Question

How can regression analysis be used to understand the cost behavior analysis for **Trex**?

Obtain the Data

Trex sells plastic/wood composite decking through its brand name, Trex Decks. Trex sells its product through **Lowe's**, and it sells most of its products in the spring and summer, when consumers install or remodel their decks.

The data set in Lab 8.4 Data.xlsx provides some weekly sample data for Trex's costs.

1. Open Excel file Lab 8.4 Data.xlsx.
2. Here is the data dictionary:

Data Dictionary

Week Beginning:	The date that the production week begins
Production:	Number of board feet of composite decking produced during the week
Total Production Cost:	Total cost of production (including all fixed and variable costs) for the production week

Analyze the Data

Analysis Task 1: Identify Which Months Correspond to the Busy Season.

To estimate the cost behavior for Trex, we will run a basic regression analysis. To prepare the data, we need to insert a "1" in the cells that correspond to the busy season and a "0" in the cells that correspond to the off-season. Because the product sells primarily in the spring and summer, Trex builds up its inventory from the beginning of the year through the first part of June.

1. Insert a new column "C", and insert the title "Busy" in cell C1. (Before you complete Steps 1–3, please see Step 4 below.)
2. Next, starting with cell C2, insert a 1 for those weeks from the beginning of the year until (and including) 6/3/2024.
3. Insert a "0" (zero) for those weeks starting 6/10/2024 until the end of the year.
4. We assume we have a huge amount of data, so we will use an Excel formula to complete Steps 1 and 2. Note that "if" statements in Excel that use dates as the logical qualifier must use a DATEVALUE function, like this: =IF(A2<DATE VALUE("6/5/2024"),1,0). This function tells Excel to look at the date in Column A, and if the date is less than 6/5/2024, put a "1" into Column C, otherwise put a 0 in column C. Add this formula to C2: =IF(A2<DATEVALUE("6/5/2024"),1,0). Then copy it down to the bottom of the data in Column C.

The first few rows of the spreadsheet should appear as shown in Lab Exhibit 8.4.1.

Lab Exhibit 8.4.1

	A	B	C	D
1	Week beginning	Productior	Busy	Cost
2	1/1/2024	90000	1	60581
3	1/8/2024	92000	1	61164
4	1/15/2024	88000	1	47439
5	1/22/2024	102000	1	63805
6	1/29/2024	98000	1	50680
7	2/5/2024	100000	1	54029
8	2/12/2024	96000	1	62131
9	2/19/2024	98000	1	48531
10	2/26/2024	101000	1	59704
11	3/4/2024	99000	1	59405
12	3/11/2024	103000	1	51458
13	3/18/2024	107000	1	61877

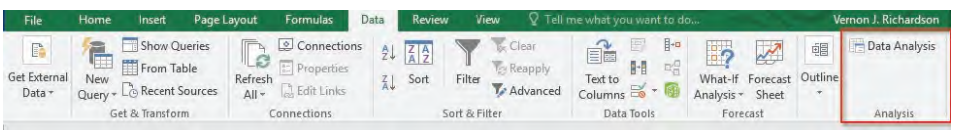
Analysis Task 2: Perform Regression Analysis to Identify Fixed and Variable Costs.

Regression analysis uses mathematics and statistics to determine which of several variables have an impact on an outcome and the size of that impact. Regression analysis helps us answer the questions: Which factors matter the most in predicting the outcome? Which factors can we ignore?

In this case, we are trying to figure out which cost drivers are associated with the outcome, or total production costs. The dependent (outcome) variable is total production costs for that production week, and the independent variable is the amount of board feet of production. We can think about this situation in terms of the algebraic equation $y = f(x)$, where y is the dependent variable and x is the independent variable. In other words, we will run an equation where total production costs (y) are a function of the amount of board feet of production (x).

Through regression analysis, we can assess the fixed cost and variable cost components of production, where the fixed cost component is the intercept and the variable cost component is the coefficient on the weekly production. Proceed as follows:

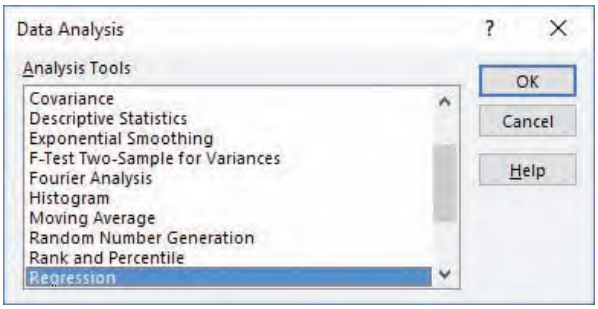
1. You will need the Excel Data Analysis ToolPak to run the regression. If you haven't installed the Data Analysis ToolPak, Appendix E provides instructions for installing it. To run the regression, make sure the Analysis ToolPak is loaded by looking at the **Data > Analysis** and seeing if the Data Analysis add-in has been installed (Lab Exhibit 8.4.2).



Lab Exhibit 8.4.2

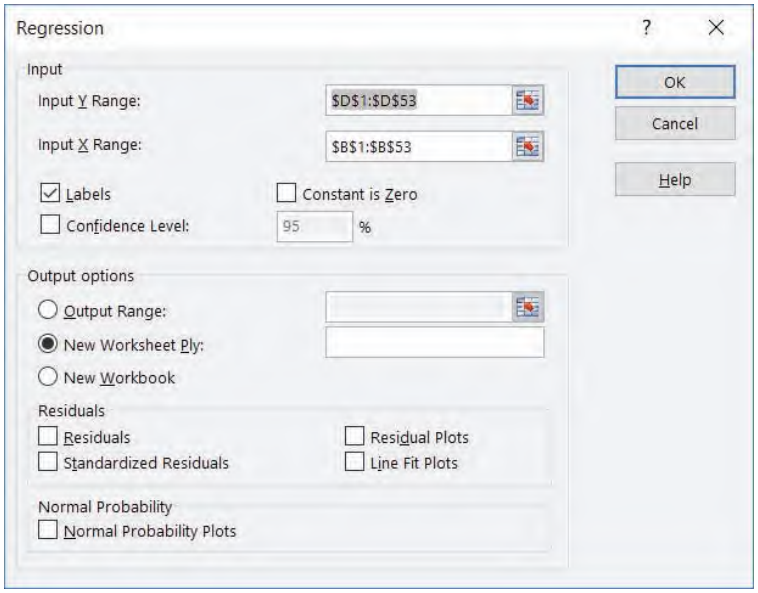
To perform the regression, select **Data> Analysis > Data Analysis**. A dialog box will open, as in Lab Exhibit 8.4.3.

Lab Exhibit 8.4.3



2. Select Regression and click OK. We need to select the dependent/outcome variable (or Y Range). Our dependent variable is Total Production Cost, so enter "$D$1:$D$53" in the Input Y Range.
3. Next, we need to select our independent variable (or X Range). We expect that production board feet (variable name: Production) is associated with total production costs, so we include production as our independent variable. Enter "$B$1:$B$53" in the Input X Range.
4. Because the data range includes data labels, click the box next to Labels. Also select "New Worksheet Ply" to get the results in a new worksheet. The completed Regression dialog box should look as shown in Lab Exhibit 8.4.4.

Lab Exhibit 8.4.4



5. Now click OK to get the regression results that are shown in Lab Exhibit 8.4.5.

There are many things to note about the regression output. The first is that the regression did better than chance at predicting total costs. Because the *Significance F* result is very small, there is an extremely small probability that the total production costs can be explained by no independent variables, which suggests that we will be able to identify the relationship between total production costs and production. The adjusted R Square at

Lab Exhibit 8.4.5

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.9439149				
5	R Square	0.890975339				
6	Adjusted R Square	0.888794846				
7	Standard Error	4325.213641				
8	Observations	52				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	1	7644094914	7.64E+09	408.6118	1.02909E-25
13	Residual	50	935373651.9	18707473		
14	Total	51	8579468566			
15						
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
17	Intercept	5259.413536	2052.824391	2.562038	0.013466	1136.194399
18	Production	0.52549949	0.025996616	20.21415	1.03E-25	0.473283749

0.889 suggests that nearly 89% of the variation in production costs can be explained by the amount of production.

Note that the *p*-value on the intercept and the coefficient on production (Rows 17 and 18) are both statistically related to total costs with a *p*-value of less than 0.05. These values suggest that we have a reasonable model for explaining the fixed and variable cost components of total costs.

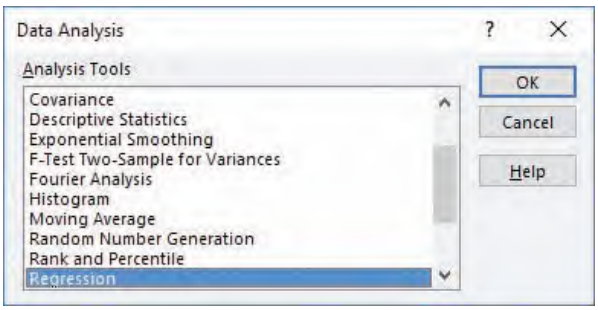
Given the regression results, we see the intercept, which is the amount of fixed costs of $5,259.41 per week. We also see that the coefficient on production suggests that the variable cost is $0.525 per board foot.

Analysis Task 2: Perform Regression Analysis to Understand the Busy Season's Impact on Costs.

We are also interested in whether the costs of the busy season also have an impact on total cost. Earlier in this lab, we inserted a dummy variable (dummy because it carries only the values of 1 and 0) in Column C for the busy season.

1. To perform our analysis, we will rerun our regression using two independent variables. Select **Data > Analysis > Data Analysis > Regression** (Lab Exhibit 8.4.6).

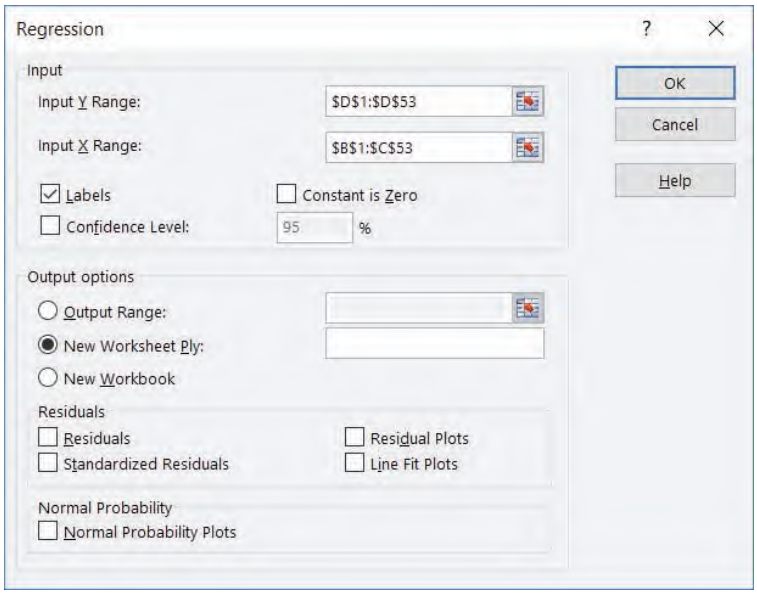
Lab Exhibit 8.4.6



2. A dialog box will open. Select Regression and click OK. We need to select the dependent variable (or Y Range). Our dependent (outcome) variable is total production, so enter "$D$1:$D$53" in the Input Y Range.
3. Next, we need to select our independent variable (or X Range). We expect that production board feet (variable name: Production) and busy season (variable name:

Busy) are associated with total production costs, so we need to include production and busy season as our independent variables. Enter "$B$1:$C$53" in the Input X Range. The completed regression dialog box should look as shown in Lab Exhibit 8.4.7.

Lab Exhibit 8.4.7



4. Click OK to get the regression results shown in Lab Exhibit 8.4.8.

Lab Exhibit 8.4.8

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.944444508
R Square	0.891975429
Adjusted R Square	0.887566263
Standard Error	4349.040265
Observations	52

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	7652675156	3.83E+09	202.3003	2.09557E-24
Residual	49	926793410.2	18914151		
Total	51	8579468566			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	3982.494805	2802.666741	1.420966	0.16166	-1649.674877
Production	0.550792858	0.045755385	12.03777	3.01E-16	0.458843969
Busy	-1431.62093	2125.551862	-0.67353	0.503775	-5703.077317

These results suggest that the busy-season costs are not associated with additional costs (because the *p*-value is greater than 0.05). In contrast, production continues to be a significant explanatory variable (with *p*-value of less than 0.05) explaining costs. The adjusted R Square at 0.888 suggests that nearly 89 percent of the variation in production costs can be explained by the amount of production.

Report the Results

We have now developed estimates for fixed and variable costs and also considered if the costs change during the busy season. To the extent that costs affect production and pricing decisions, an understanding of the costs is helpful to management in its decision-making.

Assessment

1. Take a screenshot of the regression output identifying fixed and variable costs (without the busy season variable), paste it into a Word document named "Lab 8.4 Excel Submission.docx", and label your screenshot Submission 1.
2. Answer the questions in Connect and upload your Lab 8.4 Excel Submission.docx via Connect, if assigned.

Alternate Lab 8.4 Excel

Diagnostic Analytics: Estimating Fixed and Variable Costs

Apply the same steps in Lab 8.4 to the Alt Lab 8.4 Data.xlsx data set. Thrustmaster is a fictitious company that sells game controllers; joysticks; steering wheels for PCs; and XBOX, Nintendo, and PlayStation consoles. It sells most of its products during the Christmas season.

Thrustmaster is trying to estimate its cost behavior for its basic joystick using the Alt Lab 8.4 Data.

Required

1. Use regression analysis to estimate the fixed and variable costs.
2. Use regression analysis to determine if costs are different during the busy season.
3. Use regression analysis to determine if the company is better off incorporating the busy season into its cost behavior analysis.

Data Dictionary

Week Beginning:	The date that the production week begins
Production:	Number of joysticks produced during the week
Total Production Cost:	Total cost of production (including fixed and variable costs) for the production week

Open Excel file Alt Lab 8.4 Data.xlsx. Proceed as follows:

1. Insert a busy season variable in Column C labeled "Busy" with the value of "1" during the busy season from 7/2/2022 up to and including the week of 11/26/2022. Insert the value of "0" for all other weeks.
2. Run a regression analysis similar to the one you performed in Lab 8.4, first considering only production.
3. Next, run a similar regression analysis similar to the one you performed for Lab 8.4, but now including both the busy season variable and production as independent variables.

Assessment

1. Take a screenshot of the regression output identifying fixed and variable costs, paste it into a Word document named "Alt Lab 8.4 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 8.4 Excel Submission.docx via Connect, if assigned.

Lab 8.5 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Predictive Analytics: Forecasting Future Performance of IBM

Keywords

Predictive Analysis, Forecasting, Time Series

Decision-Making Context

There are many reasons to forecast a company's future performance. For example, it may need to predict sales to determine needed manufacturing capacity, or it may be predicting cash flows to determine if it will need a loan or have sufficient cash on hand without taking a loan. In addition, an investor might be forecasting a company's earnings to determine whether the company's stock will be a good investment.

The goal of this lab is to forecast sales and earnings for **IBM** using Excel's Forecast Sheet. (*Note:* Excel's Forecast Sheet is not available on Excel for Mac computers.)

Required

1. Using Excel's Forecast Sheet, forecast future sales for IBM.
2. Using Excel's Forecast Sheet, forecast future income before extraordinary items for IBM.

Specify the Question

What will IBM's 2022, 2023, and 2024 quarterly sales and earnings be given times series (past quarterly sales and earnings from 2010 to 2021)?

Obtain the Data

Open Excel file Lab 8.5 Data.xlsx. This file contains quarterly sales and earnings for IBM from 2010 to 2021. Here is the data dictionary:

Data Dictionary

Date:	Quarter end date
IBM Sales:	Net sales for IBM during the respective quarter
IBM Income before Extraordinary Items	Income before extraordinary items for IBM during the respective quarter

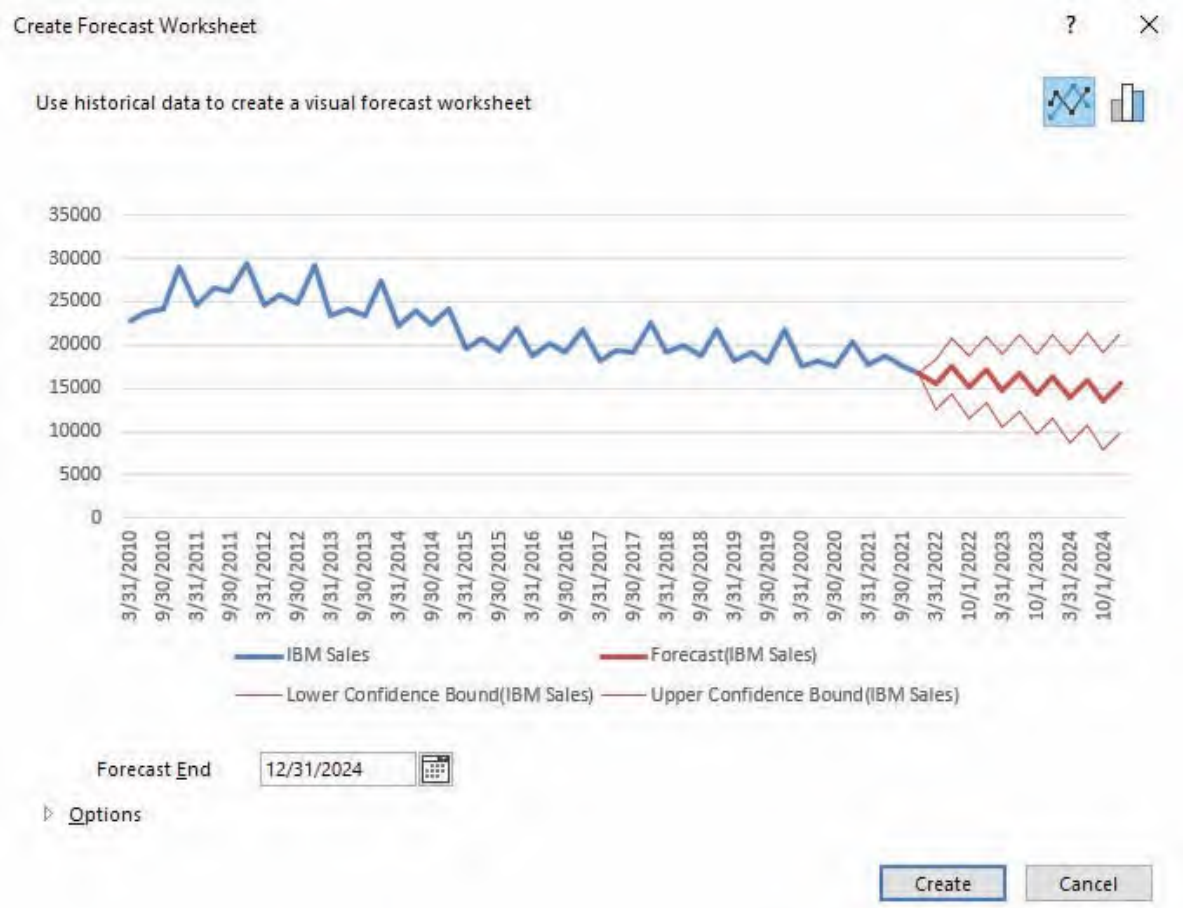


DATA VISUALIZATION

Analyze the Data

Analysis Task 1: Forecast Sales.

1. To start the forecast analysis, highlight one of the numbers in Column B (sales), and then click on **Data > Forecast > Forecast Sheet**. The dialog box in Lab Exhibit 8.5.1 will open up, beginning our forecast of IBM net sales in 2022, 2023, and 2024. The Forecast Sheet is available only for Windows, and not for Macs. If you use a Mac, you may need to use a computer lab or a virtual desktop to get access to this functionality.
2. Note the seasonality of the data with the reasonably consistent crests and troughs over the four quarters. Click on the chevron next to Options near the bottom left.
3. Consider the various options, including the dates forecast start and end, the confidence interval, the timeline range and value range, and how missing points and duplicates are handled.



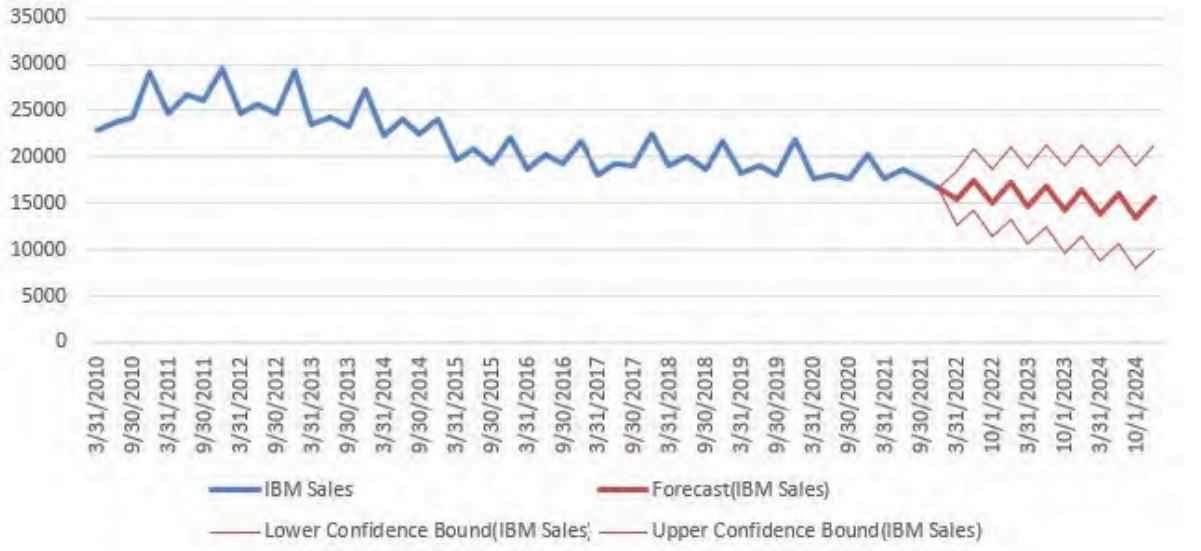
Lab Exhibit 8.5.1

4. Select "Include forecast statistics" near the bottom left and click the Create button at the bottom.
5. Note the new tab with forecast statistics showing forecasted sales for each quarter in 2022, 2023, and 2024. Also note the lower and the upper confidence bound of the forecast. The forecast sheet suggests that there is a 95% chance that the actual sales will be in the range between the lower and the upper confidence bound.
6. The forecast graph should look as shown in Lab Exhibit 8.5.2.
7. Note the widening between the confidence bounds on the graph. Calculate the difference between the lower and the upper bound, noting that, over time, the difference between the lower and upper confidence bounds gets larger because the further we get out in time, the less certainty we have of the outcome.

   - At 3/31/2022, the difference between upper and lower confidence bound is 18,441.04 − 12,569.56 = 5,871.48.
   - At 12/31/24, the difference between upper and lower confidence bound is 21,317.63 − 9,822.83 = 11,494.80.

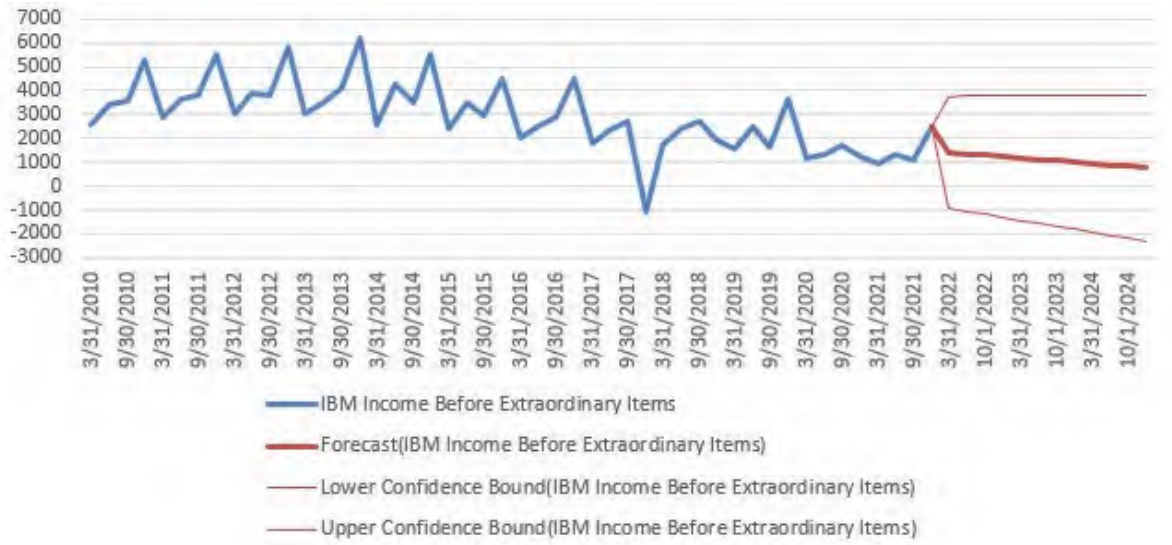
Analysis Task 2: Forecast Earnings (Income Before Extraordinary Items).

1. Perform the same analysis for Income Before Extraordinary Items with the data in Column E, repeating Steps 1–6 in Analysis Task 1. We choose income before any possible extraordinary items because predicting extraordinary items is very difficult to do.



Lab Exhibit 8.5.2

2. The graph for the forecast of future income from extraordinary items is as shown in Lab Exhibit 8.5.3.



Lab Exhibit 8.5.3

Report the Results

We have now developed a forecast for future sales and income before extraordinary items. This forecast will be helpful for management, investors, lenders, and other stakeholders.

Assessment

1. Take a screenshot of the forecast of sales, paste it into a Word document named "Lab 8.5 Excel Submission.docx", and label the screenshot Submission 1.

2. Take a screenshot of the forecast of income before extraordinary items, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 8.5 Excel Submission.docx via Connect, if assigned.

Alternate Lab 8.5 Excel

Predictive Analytics: Forecasting Future Performance of Netflix

Using the same steps in Lab 8.5 Excel, forecast future sales and earnings for **Netflix**.

Specify the Question

What will be Netflix's quarterly sales and earnings for 2022, 2023, and 2024, given time series data (past quarterly sales and earnings from 2010 to 2021)?

Required

1. Using Excel's Forecast Sheet, forecast future sales for Netflix for 2022 to 2024.
2. Using Excel's Forecast Sheet, forecast future income before extraordinary items for Netflix for 2022 to 2024.

Open Excel file Alt Lab 8.5 Data.xlsx. Here is the data dictionary:

Data Dictionary

Date:	Quarter begin date
NFLX Sales:	Net sales for Netflix during the respective quarter
NFLX Income Before Extraordinary Items:	Income before extraordinary items for Netflix during the respective quarter

Assessment

1. Take a screenshot of the forecast of sales, paste it into a Word document named "Alt Lab 8.5 Excel Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the forecast of income before extraordinary items, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 8.5 Excel Submission.docx via Connect, if assigned.

Lab 8.6 Tableau

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Predictive Analytics: Forecasting Future Performance of IBM

Keywords

Predictive Analysis, Forecasting, Time Series

Decision-Making Context

There are many reasons to forecast a company's future performance. For example, it may need to predict sales to determine needed manufacturing capacity, or it may be predicting cash flows to determine if it will need a loan or have sufficient cash on hand without taking a loan. In addition, an investor might be forecasting a company's earnings to determine whether the company's stock will be a good investment.

The goal of this lab is to forecast sales and earnings for **IBM** using Tableau.

Required

1. Using Tableau's forecasting options, forecast future sales for IBM.
2. Using Tableau's forecasting options, forecast future income before extraordinary items for IBM.

Specify the Question

What will IBM's quarterly sales and earnings be for 2022, 2023, and 2024, given time series (past quarterly sales and earnings from 2010 to 2021)?

Obtain the Data

1. The Lab 8.6 data file contains IBM's quarterly sales and earnings.
2. Open Tableau and connect to Excel file Lab 8.6 Data.xlsx.
3. Here is the data dictionary:

Data Dictionary

Date:	Quarter end date
IBM Sales:	Net sales for IBM during the respective quarter
IBM Income before Extraordinary Items:	Income before extraordinary items for IBM during the respective quarter

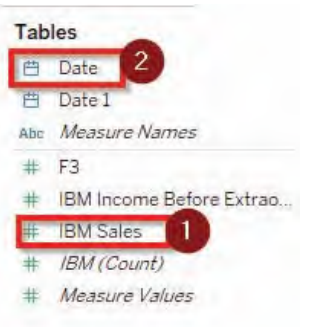
Perform the Analysis

Analysis Task 1: Forecast Sales.

Make sure the data loaded correctly. Click on Sheet 1. To start the forecast analysis, double-click IBM Sales from Tables. Then double-click Date also from Tables (as shown in Lab Exhibit 8.6.1).

Lab Exhibit 8.6.1

Tableau Software



1. Tableau should have defaulted the visualization type to a continuous line chart, but if it did not, you can change it from the Show Me tab (Lab Exhibit 8.6.2).



Lab Exhibit 8.6.2

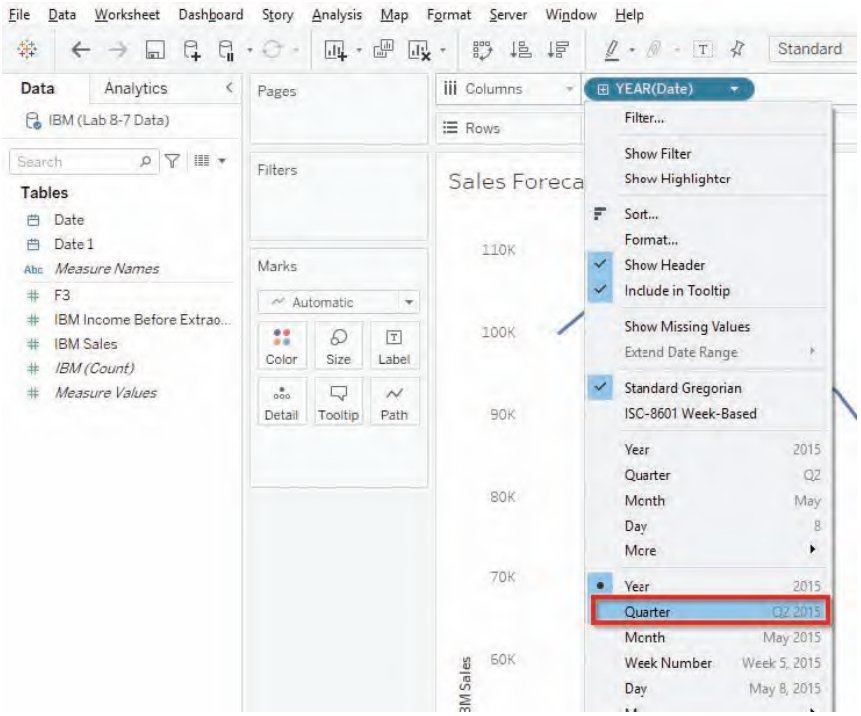
Tableau Software

Tableau interacts with dates in two ways, as either "Continuous" or "Discrete." Regardless of the way Tableau chose to display your dates when you added them to your visualization, you can change the field to the right display quite easily.

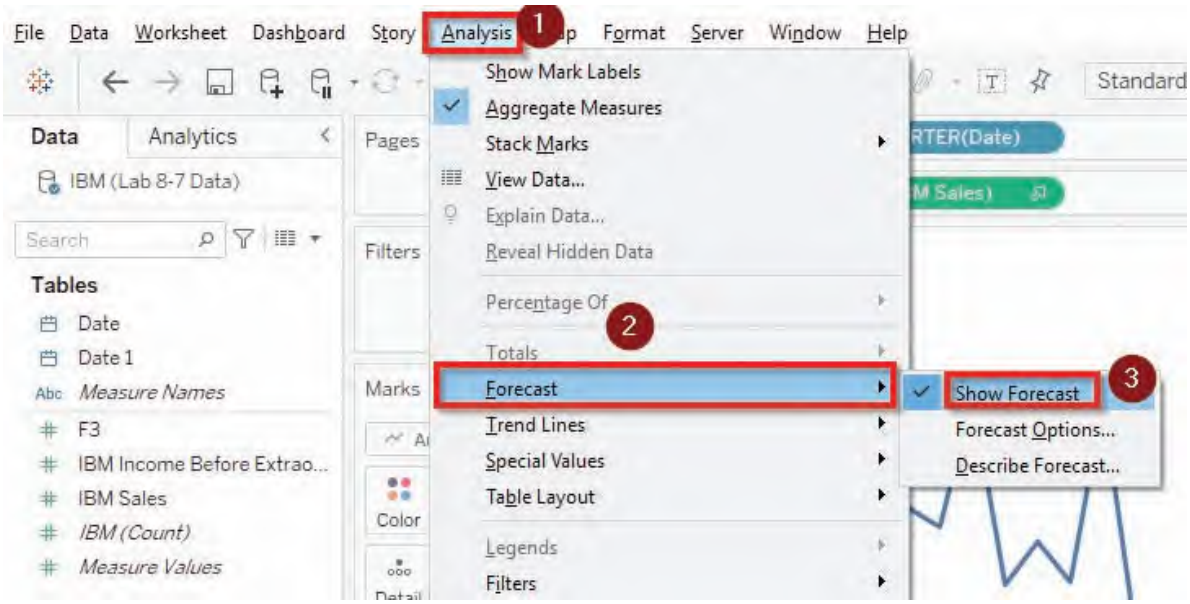
2. From the Columns shelf, right-click the YEAR(Quarter) pill, and select Quarter Q2 2015 (Lab Exhibit 8.6.3).

Lab Exhibit 8.6.3

Tableau Software



3. To expand this line chart to show a Forecast, select the **Analysis Tab > Forecast > Show Forecast** (Lab Exhibit 8.6.4). You can now select a variety of options to refine this forecast.



Lab Exhibit 8.6.4

Tableau Software

4. From the Marks shelf, change the default mark to Circle (Lab Exhibit 8.6.5).



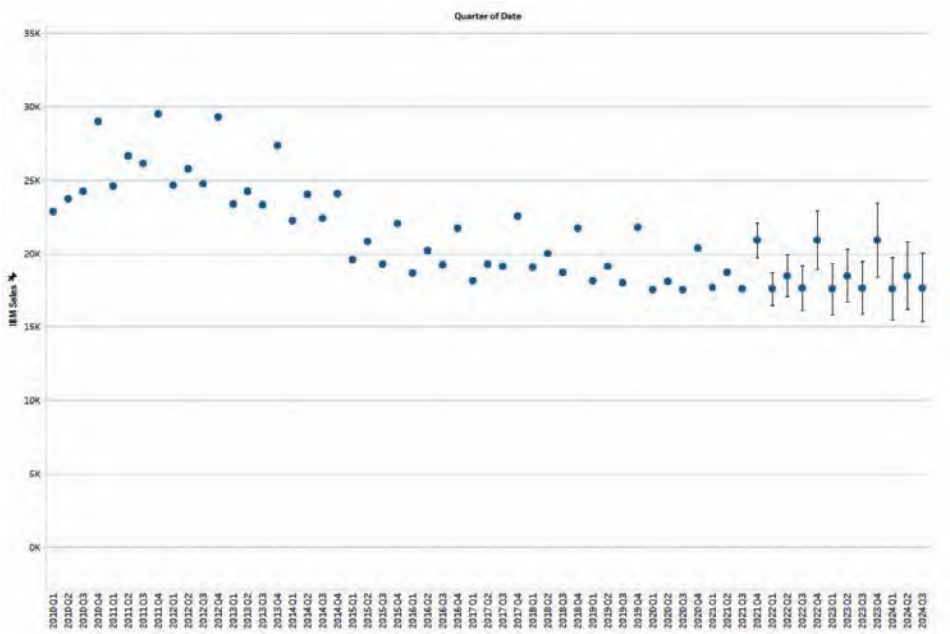
Lab Exhibit 8.6.5

Tableau Software

The line chart has changed, and Tableau has provided whiskers around each of the forecasted data points (Lab Exhibit 8.6.6).



DATA VISUALIZATION



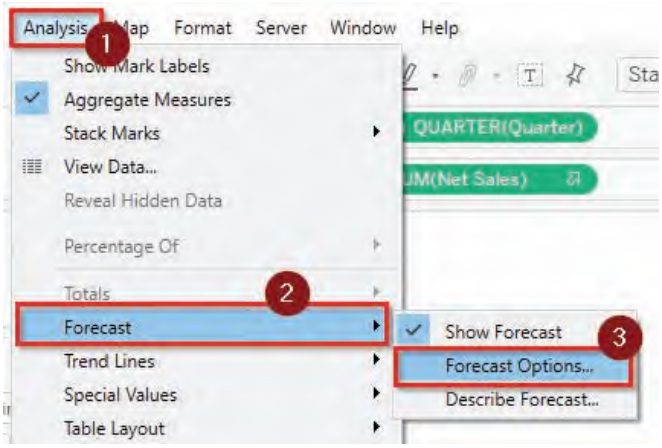
Lab Exhibit 8.6.6

Tableau Software

5. You can also adjust the amount of time you are forecasting. From the Analysis Tab, select **Forecast > Forecast Options** (Lab Exhibit 8.6.7).

Lab Exhibit 8.6.7

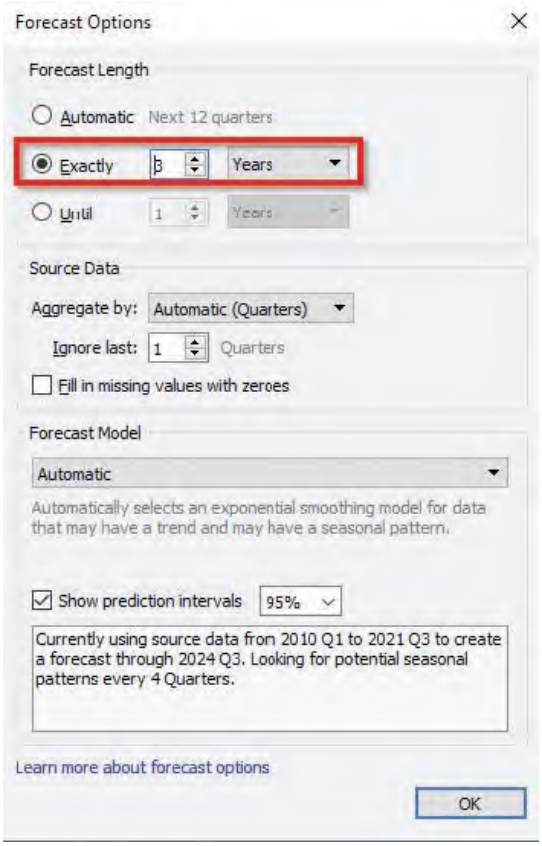
Tableau Software



6. Adjust the Forecast Length to Exactly 3 Years, and then click OK (Lab Exhibit 8.6.8).

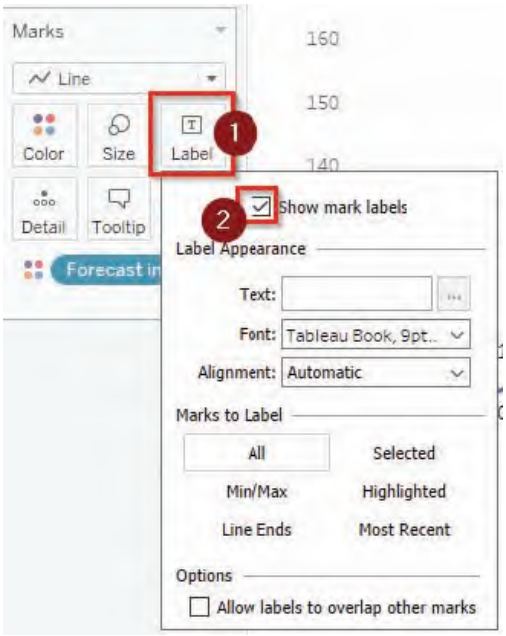
Lab Exhibit 8.6.8

Tableau Software



As you make the change, you can see how the chart adjusted to provide more data. You can also see that the whiskers around the dates extend as you reach further into the future, due to increasing uncertainty about the outcome.

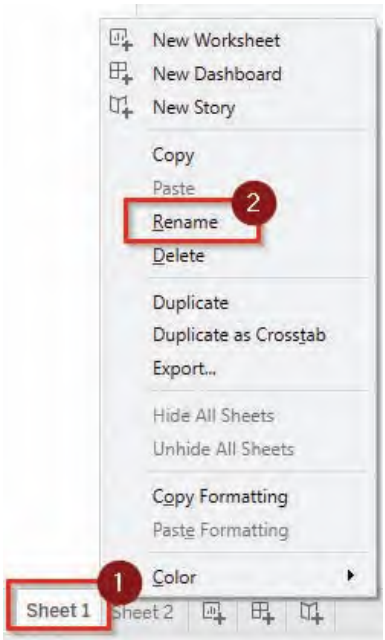
7. Notice that the labels indicating the month and product demand (either actual or estimated) do not appear by default. To add the labels, click Label in the Marks shelf and add a checkmark next to Show mark labels (Lab Exhibit 8.6.9).



Lab Exhibit 8.6.9

Tableau Software

8. Rename this sheet to "Sales Forecast with Box Plots". Right-click on Sheet 1 and select Rename (Lab Exhibit 8.6.10).



Lab Exhibit 8.6.10

Tableau Software

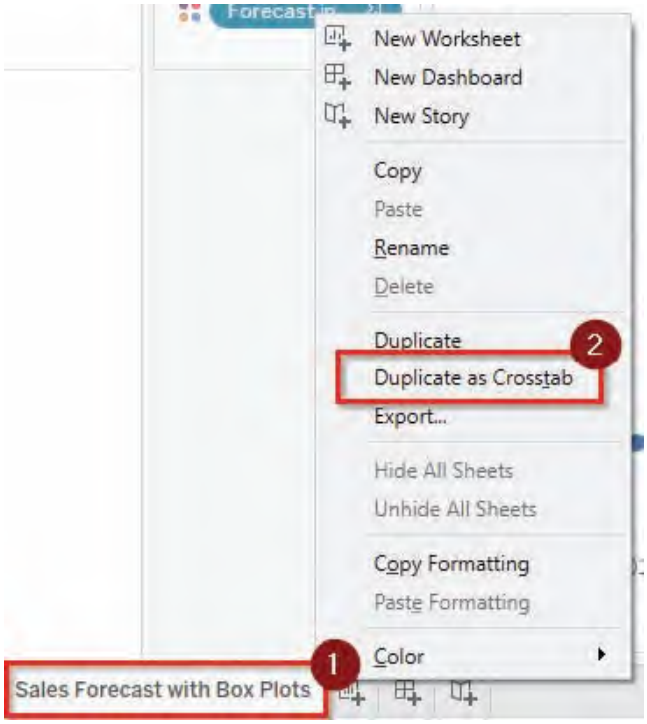
9. Tableau does not default to providing the values for the upper and lower prediction levels the way Excel does. Remember, Tableau is based primarily on visualizing data.

However, when forecasting, it can be useful to see not only the marks for the predicted values but also the upper and lower prediction levels.

To see these levels, we need to duplicate our current sheet as a cross-tab. Right-click Sales Forecast with Box Plots and select Duplicate as Crosstab (Lab Exhibit 8.6.11).

Lab Exhibit 8.6.11

Tableau Software



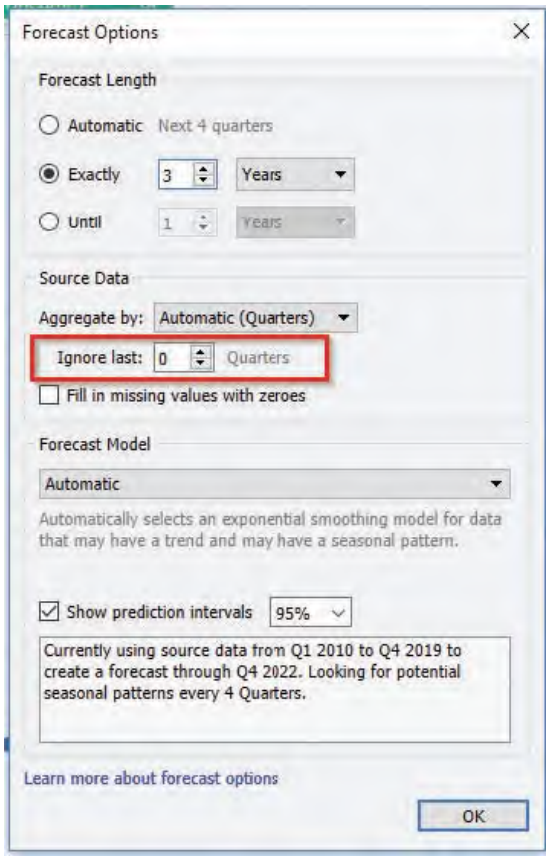
10. You will see a new sheet titled Sales Forecast with Box Plots (2). Rename this sheet to "Sales Forecast Crosstab Data."
11. Click into your new Sales Forecast with Crosstab Data sheet to view the cross-tab data. If you scroll to the right, you will see data for the upper and lower bounds.

Analysis Task 2: Forecast Earnings (Income Before Extraordinary Items).

1. Create a new sheet and perform the same analysis for Income before extraordinary items, repeating Steps 1–11 from Analysis Task 1 with the new data (the only change is using the Income measure instead of Net Sales). We choose income before any possible extraordinary items because predicting extraordinary items is very difficult to do.
   If you completed Lab 8.5 before this lab, you will notice that the forecast is vastly different. This is because Tableau has defaulted to leaving out the last quarter from its analysis.
2. Select **Analysis Tab > Forecast > Forecast Options** and adjust the default from 1 to 0 in Ignore Last Quarters. Then click OK (Lab Exhibit 8.6.12).

Report the Results

We have now developed a forecast for future sales and income before extra items. This forecast will be helpful for management, investors, lenders, and other stakeholders.



Lab Exhibit 8.6.12

Tableau Software

Assessment

1. Take a screenshot of the forecast of sales, paste it into a Word document named "Lab 8.6 Tableau Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the forecast of income, paste it into the same Word document, and label the screenshot Submission 3.
3. Answer the questions in Connect and upload your Lab 8.6 Tableau Submission.docx via Connect, if assigned.

Alternate Lab 8.6 Tableau

Predictive Analytics: Forecasting Future Performance of Netflix

Using Lab 8.6 Tableau as a guide, forecast future sales and income before extraordinary items for **Netflix**.

Required

1. Using Tableau's forecasting options, forecast future sales for Netflix.
2. Add a duplicate of the sheet to show the cross-tab data.
3. Using Tableau's Forecasting options, forecast future income before extraordinary items for Netflix. (Don't forget to adjust the amount of quarters ignored to 0.)

4. Add a duplicate of the sheet to show the cross-tab data.

Open Tableau and connect to Excel file Alt Lab 8.6 Data.xlsx.

Data Dictionary

Date:	Quarter begin date
NFLX Sales:	Net sales for Netflix during the respective quarter
NFLX Income Before Extraordinary Items:	Income before extraordinary items for Netflix during the respective quarter

Assessment

1. Take a screenshot of the forecast of sales, paste it into a Word document named "Alt Lab 8.6 Tableau Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the forecast of income, paste it into the same Word document, and label the screenshot Submission 3.
3. Answer the questions in Connect and upload your Alt Lab 8.6 Tableau Submission.docx via Connect, if assigned.

Lab 8.7 Power BI

Predictive Analytics: Forecasting Future Performance of IBM

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Predictive Analysis, Forecasting, Time Series

Decision-Making Context

There are many reasons to forecast a company's future performance. For example, it may need to predict sales to determine needed manufacturing capacity, or it may be predicting cash flows to determine if it will need a loan or have sufficient cash on hand without taking a loan. In addition, an investor might be forecasting a company's earnings to determine whether the company's stock will be a good investment.

The goal of this lab is to forecast sales and earnings for **IBM** using Power BI.

Required

1. Using Power BI's forecasting options, forecast future sales for IBM.
2. Using Power BI's forecasting options, forecast future income before extraordinary items for IBM.

Specify the Question

What will IBM's quarterly sales and earnings be for 2022, 2023, and 2024, given time series data (past quarterly sales and earnings from 2010 to 2021)?

Obtain the Data

Looking at IBM's financial statements, we accumulate the values of quarterly sales and earnings from 2010 to 2021.

1. Open Power BI and connect to Excel file Lab 8.7 Data.xlsx. Click the IBM sheet and click Load. Here is the data dictionary:

Data Dictionary

Date:	Quarter end date
IBM Sales:	Net sales for IBM during the respective quarter
IBM Income before Extraordinary Items:	Income before extraordinary items for IBM during the respective quarter

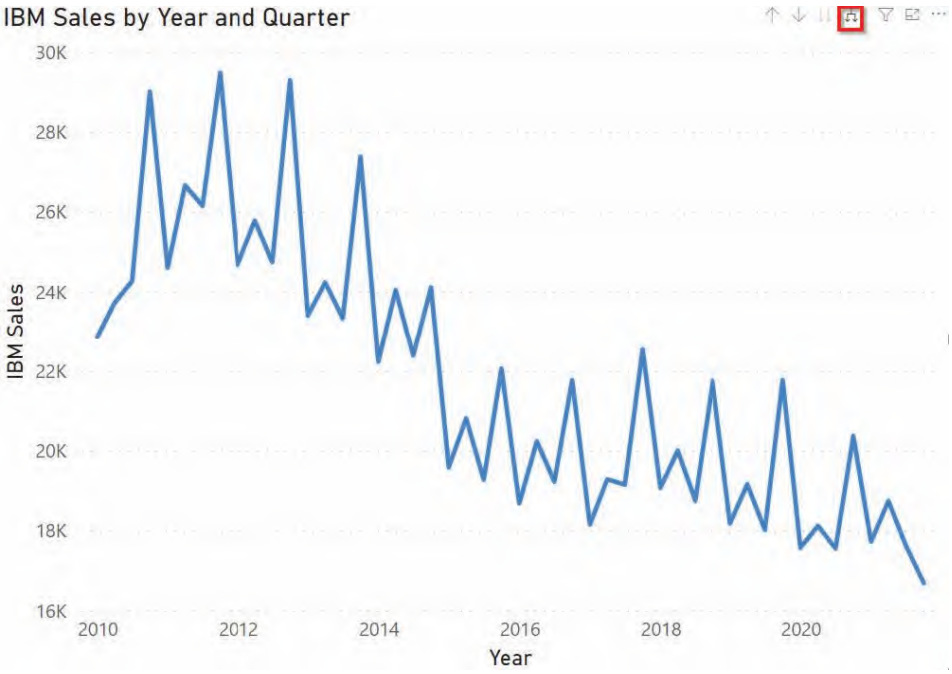
Perform the Analysis

Analysis Task 1: Forecast Sales.


DATA VISUALIZATION

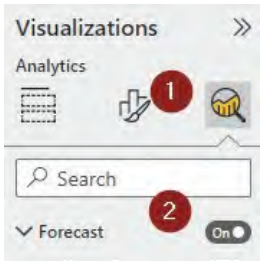
1. To start the forecast analysis, click the Line chart button in the Visualizations pane. Check the boxes next to IBM Sales and Date to add them to your visual as Values and Axis, respectively.
2. To expand the dates to show Quarters instead of Years, click Expand all down one level in the hierarchy (upside-down fork button in the top-right corner of the visual), as shown in Lab Exhibit 8.7.1.

Lab Exhibit 8.7.1



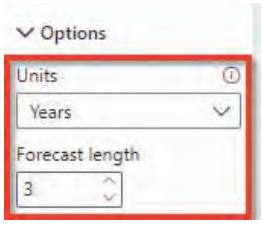
3. To expand this line chart to show a Forecast, click the Analytics magnifying glass button in the Visualizations pane. Click on the Forecast toggle to turn the forecasting function "on" (as shown in Lab Exhibit 8.7.2).

Lab Exhibit 8.7.2



4. Adjust the Forecast length to Exactly 3 Years (Lab Exhibit 8.7.3). Click Apply.

Lab Exhibit 8.7.3



Analysis Task 2: Forecast Earnings (Income Before Extraordinary Items).

1. Create a new line chart and do the same analysis for Income before Extraordinary Items, repeating Steps 1–4 from Analysis Task 1 with the new data. We choose income before any possible extraordinary items because extraordinary items are rare and hard to predict.

Report the Results

We have now developed a forecast for future sales and income before extra items. This forecast will be helpful for management, investors, lenders, and other stakeholders.

Assessment

1. Take a screenshot of the forecast of sales, paste it into a Word document named "Lab 8.7 Power BI Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the forecast of income, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 8.7 Power BI Excel Submission.docx via Connect, if assigned.

Alternate Lab 8.7 Power BI

Predictive Analytics: Forecasting Future Performance of Netflix

Using Lab 8.7 Power BI as a guide, forecast future sales and income before extraordinary items for **Netflix**.

Required

1. Using Power BI's forecasting options, forecast future sales for Netflix.
2. Using Power BI's forecasting options, forecast future income before extraordinary items for Netflix.

   Open Power BI and connect to Excel file Alt Lab 8.7 Data.xlsx. Here is the data dictionary:

Data Dictionary

Date:	Quarter begin date
NFLX Sales:	Net sales for Netflix during the respective quarter
NFLX Income Before Extraordinary Items	Income before extraordinary items for Netflix during the respective quarter

Assessment

1. Take a screenshot of the forecast of sales, paste it into a Word document named "Alt Lab 8.7 Power BI Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of the forecast of income, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 8.7 Power BI Submission.docx via Connect, if assigned.

Lab 8.8 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Prescriptive Analytics: Using Goal-Seek Analysis to Determine the Break-Even Point

Keywords

What-if Solver, Break-even, Goal Seek

Decision-Making Context

Managers use historical data to predict the costs and prices of products. What-if analysis helps managers estimate the company's break-even point. The estimated break-even point sets the threshold for the positive net income and the goal for minimum sales.

In this lab, we use What-if Analysis in Excel to estimate break-even point.

Required

Use What-if Analysis in Excel to predict the break-even point.

Specify the Question

What is the break-even point given costs and sales prices?

Obtain the Data

Cajun Cooking is a food consumer packaged goods company that is considering selling a new Creole seasoning packet. The costs of ingredients for food production vary by season. In terms of costs and sales prices, the company's management wants to estimate the break-even point for Creole Seasoning Packet in each month to set up minimum sales targets.

1. Open Excel file Lab 8.8 Data.xlsx.
2. Browse the spreadsheet quickly to make sure there aren't any obvious errors in the Excel file.

Data Dictionary

Price:	The unit price of the Creole Seasoning Packet
Unit sales:	The expected unit sales of the Creole Seasoning Packet per month
Unit cost:	The variable costs per Creole Seasoning Packet
Fixed cost:	The fixed costs of each month
Total variable cost:	Total variable costs of the Creole Seasoning Packet per month
Total cost:	The total fixed and variable costs of the Creole Seasoning Packet per month
Net income:	The net income of the Creole Seasoning Packet per month

Analyze the Data

Analysis Task 1: Calculate Costs and Revenues.

1. Calculate the total variable costs of the Creole Seasoning Packet in January. Variable costs = Unit sales * Unit cost (Lab Exhibit 8.8.1).

Lab Exhibit 8.8.1

	A	B	C	D	E	F
1		Jan	Feb	March	April	May
2	Price	7.00	7.00	7.00	7.00	7.00
3	Unit of sales	300.00	300.00	300.00	300.00	300.00
4	Unit cost	4.50	3.80	3.80	3.80	3.50
5	Fixed cost	1000.00	1000.00	1000.00	1000.00	1000.00
6	Total variable cost	=B3*B4				
7	Total cost					
8	Sales revenue					
9	Net income					

2. Calculate the total cost of the Creole Seasoning Packet in January. Total cost = Variable cost + Fixed cost (Lab Exhibit 8.8.2).

Lab Exhibit 8.8.2

	A	B	C	D	E	F	G
1		Jan	Feb	March	April	May	June
2	Price	7.00	7.00	7.00	7.00	7.00	7.00
3	Unit of sales	300.00	300.00	300.00	300.00	300.00	300.00
4	Unit cost	4.50	3.80	3.80	3.80	3.50	3.50
5	Fixed cost	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00
6	Total variable cost	1350.00					
7	Total cost	=B5+B6					
8	Sales revenue						
9	Net income						

3. Calculate sales revenue of the Creole Seasoning Packet in January. Sales revenue = Price * Unit sales (Lab Exhibit 8.8.3).

Lab Exhibit 8.8.3

	A	B	C	D	E	F
1		Jan	Feb	March	April	May
2	Price	7.00	7.00	7.00	7.00	7.00
3	Unit of sales	300.00	300.00	300.00	300.00	300.00
4	Unit cost	4.50	3.80	3.80	3.80	3.50
5	Fixed cost	1000.00	1000.00	1000.00	1000.00	1000.00
6	Total variable cost	1350.00				
7	Total cost	2350.00				
8	Sales revenue	=B2*B3				
9	Net income					

4. Calculate net income of the Creole Seasoning Packet in January. Net income = Sales revenue – Total cost (Lab Exhibit 8.8.4).

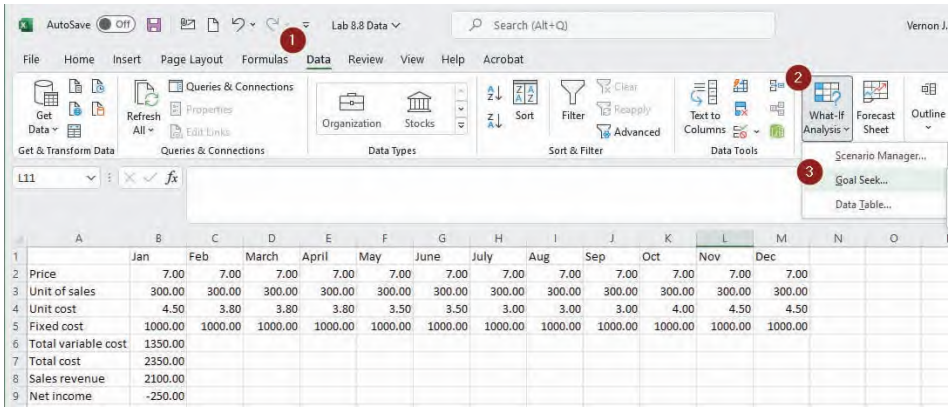
Lab Exhibit 8.8.4

	A	B	C	D	E
1		Jan	Feb	March	April
2	Price	7.00	7.00	7.00	7.00
3	Unit of sales	300.00	300.00	300.00	300.00
4	Unit cost	4.50	3.80	3.80	3.80
5	Fixed cost	1000.00	1000.00	1000.00	1000.00
6	Total variable cost	1350.00			
7	Total cost	2350.00			
8	Sales revenue	2100.00			
9	Net income	=B8-B7			

Analysis Task 2: Calculate the Break-Even Point.

1. In cell B9, estimate the break-even point for revenue and total costs using What-if solver. Go to Data in the ribbon, and in the Forecast section, click What-If Analysis. In the submenu, click Goal Seek (Lab Exhibit 8.8.5).

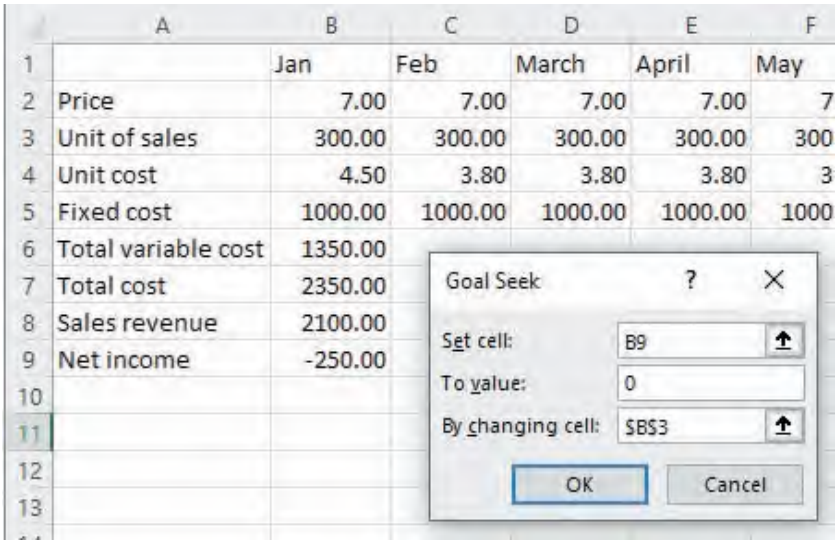
Lab Exhibit 8.8.5



	A	Jan	Feb	March	April	May	June	July	Aug	Sep	Oct	Nov	Dec
2	Price	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00
3	Unit of sales	300.00	300.00	300.00	300.00	300.00	300.00	300.00	300.00	300.00	300.00	300.00	300.00
4	Unit cost	4.50	3.80	3.80	3.80	3.50	3.50	3.00	3.00	3.00	4.00	4.50	4.50
5	Fixed cost	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00
6	Total variable cost	1350.00											
7	Total cost	2350.00											
8	Sales revenue	2100.00											
9	Net income	-250.00											

2. Estimate the unit sales when net income equals to zero. In the Goal Seek window, select net income in January (data under a condition) in the box next to Set cell. Then type "0" (the condition) in the box next to To value, select unit of sales in January in the box next to By changing cell, and click OK (Lab Exhibit 8.8.6).

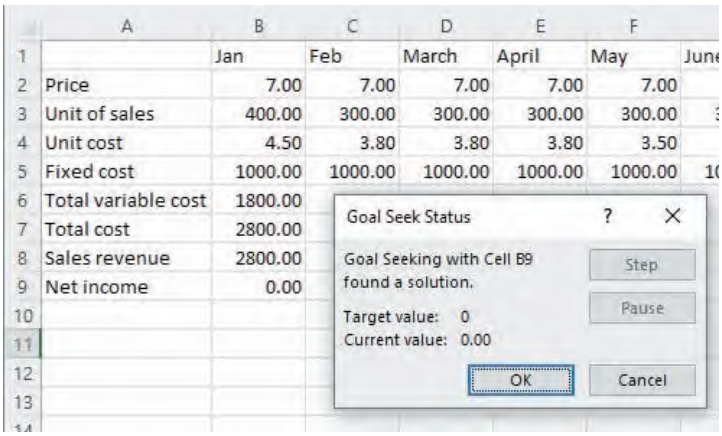
Lab Exhibit 8.8.6



	A	B	C	D	E	F
1		Jan	Feb	March	April	May
2	Price	7.00	7.00	7.00	7.00	7
3	Unit of sales	300.00	300.00	300.00	300.00	300
4	Unit cost	4.50	3.80	3.80	3.80	3
5	Fixed cost	1000.00	1000.00	1000.00	1000.00	1000
6	Total variable cost	1350.00				
7	Total cost	2350.00				
8	Sales revenue	2100.00				
9	Net income	-250.00				
10						
11						
12						
13						

3. Click OK in the Goal Seek Status window. Once you press OK, notice that the number of units in cell B3 changes to 400.00, which is the estimated unit sales in January that will result in net income equal to zero—that is, the break-even point (Lab Exhibit 8.8.7).

Lab Exhibit 8.8.7



	A	Jan	Feb	March	April	May	June
2	Price	7.00	7.00	7.00	7.00	7.00	
3	Unit of sales	400.00	300.00	300.00	300.00	300.00	
4	Unit cost	4.50	3.80	3.80	3.80	3.50	
5	Fixed cost	1000.00	1000.00	1000.00	1000.00	1000.00	1(
6	Total variable cost	1800.00					
7	Total cost	2800.00					
8	Sales revenue	2800.00					
9	Net income	0.00					
10							
11							
12							
13							

4. To identify the number of units sold in order to result in a net income equal to zero, or the break-even point, follow Steps 1 to 4 in Analysis Task 1 for the months February through December to predict unit sales of the Creole Seasoning Packet at the break-even point (Lab Exhibit 8.8.8).

Lab Exhibit 8.8.8

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Jan	Feb	March	April	May	June	July	Aug	Sep	Oct	Nov	Dec
2	Price	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00
3	Unit of sales	400.00	312.50	312.50	312.50	285.71	285.71	250.00	250.00	250.00	333.33	400.00	400.00
4	Unit cost	4.50	3.80	3.80	3.80	3.50	3.50	3.00	3.00	3.00	4.00	4.50	4.50
5	Fixed cost	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00
6	Total variable cost	1800.00	1187.50	1187.50	1187.50	1000.00	1000.00	750.00	750.00	750.00	1333.33	1800.00	1800.00
7	Total cost	2800.00	2187.50	2187.50	2187.50	2000.00	2000.00	1750.00	1750.00	1750.00	2333.33	2800.00	2800.00
8	Sales revenue	2800.00	2187.50	2187.50	2187.50	2000.00	2000.00	1750.00	1750.00	1750.00	2333.33	2800.00	2800.00
9	Net income	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Report the Results

Using what-if analysis, managers are able to predict the level of unit sales for the break-even of net income. Managers can set the level of unit sales as the threshold and adjust sales strategies to meet that threshold.

Assessment

1. Take a screenshot of the table in Step 4 of Analysis Task 2, paste it into a Word document named "Lab 8.8 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Lab 8.8 Excel Submission.docx via Connect, if assigned.

Alternate Lab 8.8 Excel

Prescriptive Analytics: Using Goal Seek Analysis to Determine the Break-Even Point

The data dictionary for Alt Lab 8.8.xlsx data set is the same as Lab 8.8.xlsx data set.

Apply the same steps in Lab 8.8 to the Alt Lab 8.8.xlsx data set, which includes sales information of products A, B, C, D, E, F, and G in 2022.

Required

Estimate the break-even point of unit sales for each product, where net income equals to zero.

Assessment

1. Take a screenshot of the table in Step 4 of Analysis Task 4, paste it into a Word document named "Alt Lab 8.8 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 8.8 Excel Submission.docx via Connect, if assigned.

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Chapter 9

Financial Analytics

A Look Back

- Accounting is a systematic process of measuring, recording, and communicating the financial information that decision-makers need to make decisions.
- Chapter 8 investigated how business analysts can answer accounting questions by identifying accounting data sources, applying analytics techniques, and reporting results to decision-makers.

A Look at This Chapter

- Finance is the management of money (or financial capital) by investing, borrowing, lending, budgeting, saving, and/or forecasting.
- Chapter 9 explains how business analysts can answer finance questions by identifying finance data sources, applying analytics techniques, and reporting results to decision-makers.

A Look Ahead

- Chapter 10 applies the SOAR model to operational analytics, including human resources (HR), information technology operations, and supply chain analytics.
- The chapter also provides examples of operational questions addressable with business analytics, examines the sources of operational data, and describes the key types of analytics used. In addition, it shows how the results of operational analytics are reported.

The Sensitivity of Husqvarna Group's Profits to Factors It May Not Control

Managers want to know how a company's financial-performance outcomes, such as operating income, are likely to respond to changes in the world that they generally cannot control, such as interest rates, the cost of raw materials, and fluctuating exchange rates. *Sensitivity* analysis, an important component of financial analytics, helps a company evaluate outcomes based on uncertainty regarding factors it does not control. Consider the following sensitivity analysis, which shows the expected results on **Husqvarna Group**'s operating income. Husqvarna is a Swedish company that manufacturers outdoor power products such as chainsaws, garden tractors, and mowers.

In its annual report, Husqvarna reports the following sensitivity analysis:



[1] Excluding hedge effects. SEK –225m refer to effect of transactions and SEK 35m to translation effects.
[2] Excluding hedge effects. SEK 400m refer to effect of transactions and SEK 100m to translation effects.
[3] Excluding hedge effects. SEK 590m refer to effect of transactions and SEK 175m to translation effects.
All others refer to the other currencies including USD and EUR.

According to this analysis, a 10 percent decrease in the cost of steel (as a raw material) will increase profits by 160 million SEK (Swedish kronor). A 1 percent decrease in interest rates will increase profits by 25 million Swedish kronor.

How will Husqvarna plan for each of these possibilities? How should it protect itself and its customers against changes in currency, the cost of raw materials, and interest rates? Which of these changes in currency, raw materials, or interest rates are expected to occur, and how will they impact Husqvarna's overall profitability and cash flows?

Source: Sensitivity Analysis (as reported in Husqvarna Annual Report). Available online at https://www.husqvarnagroup.com/sites/default/files/pr/202003118399.1.pdf. Accessed April 2020.

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

- **LO 9.1** Define *finance* and its three primary branches: corporate finance, investments, and financial markets and institutions.
- **LO 9.2** Identify representative finance questions addressable by business analytics.
- **LO 9.3** Enumerate the sources of financial data.
- **LO 9.4** Describe and provide examples of descriptive analytics used in financial analysis.
- **LO 9.5** Describe the types of finance questions addressed by diagnostic analytics.
- **LO 9.6** Describe the types of finance questions addressed by predictive analytics.
- **LO 9.7** Describe the basics of prescriptive financial analytics.
- **LO 9.8** Identify methods of reporting financial analytics results.

LABS FOR CHAPTER 9

- **LAB 9.1** **Excel:** Descriptive and Diagnostic Analytics: Calculating Returns to Investments Using the Sharpe Ratio
- **LAB 9.2** **Excel:** Diagnostic Analytics: Applying DuPont Analysis of Financial Performance
- **LAB 9.3** **Excel:** Predictive Analytics: Evaluating Loan Acceptance
- **LAB 9.4** **Excel:** Predictive Analytics: Predicting Bankruptcy
- **LAB 9.5** **Excel:** Prescriptive Analytics: Evaluating Investments Using NPV
- **LAB 9.6** **Excel:** Prescriptive Analytics: Evaluating Investments Using IRR

LO 9.1

Define *finance* and its three primary branches: corporate finance, investments, and financial markets and institutions.

9.1 THE ROLE OF FINANCE IN BUSINESS

As we saw in Chapter 8, accounting is the financial language of business. In contrast, **finance** is the management of money (or financial capital) by investing, borrowing, lending, budgeting, saving, and/or forecasting. **Financial capital** (or simply **capital**) refers to the economic resources that companies use to operate their business. Because running a business requires the careful management of financial capital, companies typically have a finance department that manages their monetary resources.

The Three Primary Branches of Finance

The three primary branches of finance are corporate finance, investments, and financial markets and institutions. Although each branch performs specific duties, they all work to manage the company's money. Let's look at each branch in a bit more detail.

Corporate Finance

Corporate finance is the business function that manages funding sources, capital structure, and investment decisions with the goal of increasing the company's value for its shareholders (or owners). It is generally performed internally rather than externally. Those who work in corporate finance budget and manage working capital and perform financial analysis. **Working capital** is the available capital, such as cash and other resources, that a company can readily use in its day-to-day operations.

Financial analysis is usually performed to analyze a company's **liquidity** (the availability of cash to meet company obligations), **solvency** (financial ability to pay company debts when due), and investment potential. It may also be used internally to allocate financial capital to different subsidiaries, divisions, departments, operations, or processes or to invest in physical assets (such as buildings or new equipment) according to their expected return or profitability.

Investments

The second branch of finance is **investments**, defined as the acquisition and disposition of assets with the objective of generating income or profit. Pension funds, mutual funds, and other companies are all looking for opportunities to invest in financial assets, such as equity investments in other companies (for example, shares of stock); U.S. Treasury bonds; corporate bonds; and commodities such as oil, wheat, gold, and pork bellies. Those who work in the investments area evaluate the likely returns given the risk of each potential investment opportunity. Based on their analysis, they recommend the best investment for the company at a point in time.

The investments function is performed both internally (looking for the best investments for a specific company) as well as externally by outside investors such as hedge funds, mutual funds, and pension funds.

Financial Markets and Institutions

Financial markets and institutions are institutions outside of the company. Companies use these institutions to raise the funds they need to start a business, buy assets, engage in mergers and acquisitions, and participate in other activities that will help them earn a profit.

Financial markets are the markets for selling various types of investments, such as stocks and bonds. They include the New York Stock Exchange, NASDAQ, and Chicago Board Options Exchange (CBOE).

Financial institutions are institutions that work in the financial markets, such as banks (for example, **Bank of America** and **Wells Fargo**), investment companies (for example,

pension investment firms such as **CalPERS**), and brokerages (such as **Goldman Sachs** and **Merrill Lynch**). Financial institutions generally act as intermediaries, or brokers, that help companies and individuals needing capital (money) find those willing to lend the capital.



FinancePix/Alamy Stock Photo

DATA ANALYTICS AND YOUR CAREER

The median annual salary for a financial analyst is $83,660. In 2020, there were 492,100 financial analysts, and the expected growth rate in the number of positions is 6% annually. What do financial analysts do? Basically, they work with the SOAR model every day, specifying finance questions and using financial data to answer those questions to inform financial decision-making.[1]

PROGRESS CHECK

1. Which branch of finance includes banks that make car loans? Explain.
2. Which branch of finance would manage a financial portfolio for the retirement fund of school teachers in New Jersey? Explain.
3. Which branch of finance would help the 7-Eleven chain of convenience stores decide whether to build more stores in China, Mexico, or Canada?

9.2 SPECIFYING THE FINANCE QUESTION

LO 9.2

Identify representative finance questions addressable by business analytics.

Specifying the finance question is the first step of the SOAR model in financial analytics. Each branch of finance has different questions that are potentially answerable with data. Exhibit 9.1 provides examples of financial questions according to analytics type (descriptive, diagnostic, predictive, and prescriptive) and the three main branches of finance.

[1] *Occupational Outlook Handbook for a Financial Analyst,* U.S. Bureau of Labor Statistics. Available online at https://www.bls.gov/ooh/business-and-financial/financial-analysts.htm. Accessed October 2021.

Financial Data Analytics Type	Corporate Finance	Investments	Financial Markets and Institutions
Descriptive: What happened? What is happening?	What was the company's return on investment for investments in the international and domestic divisions last year?	What is the return on investing in the stock market or the corporate bond market over the previous five years?	What is the sum of real-estate loans made by banks in Illinois during the past six months?
Diagnostic: Why did it happen? What are the causes of past results? Why are the results different than expectations?	Why is the company's performance different from its budgeted performance?	Is there a relationship between the returns earned and the risk borne for financial instruments?	Why did gold prices increase while silver prices declined?
Predictive: Will it happen in the future? What is the probability something will happen? Can we forecast what will happen?	What is the level of expected cash flows in the next month, quarter, and year to help prepare future budgets?	What is the chance that a company will go bankrupt? Will investors still be willing to invest in that company?	Will a potential borrower repay a loan (based on credit score, payment history, existing debt, and so on)?
Prescriptive: What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints?	Which assets should **John Deere** invest in: a new piece of manufacturing equipment or new supply chain software? Which has the greater return?	What are the expected returns to different financial instruments, given the expected risks?	What will happen to the profitability of banks in the United States if interest rates increase by 1 percent?

Exhibit 9.1
Financial Data Analytics: The Different Types of Finance Questions in the Three Main Branches of Finance

PROGRESS CHECK

4. Why is it important to predict the bankruptcy risk of a company you might invest in?
5. Give two examples of investments-related finance questions that management might ask (and that business analysts might answer).
6. Give two examples of financial questions associated with predictive analytics.

LO 9.3

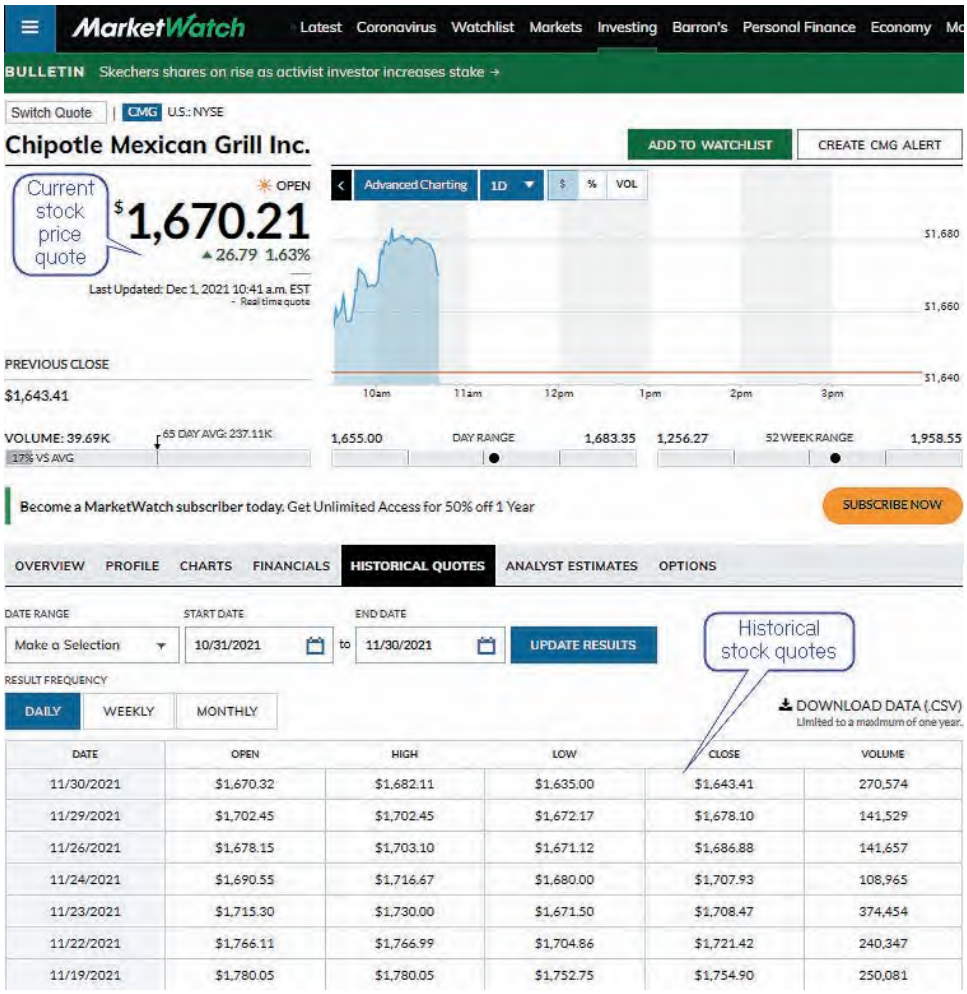
Enumerate the sources of financial data.

9.3 OBTAIN THE DATA: FINANCIAL DATA SOURCES

There are three primary sources of data typically available for financial analytics: stock-return data, summarized financial data, and financial statement data. We examine each of these financial data sources in the following sections.

Stock-Return Data

Current and historical stock-price data are important for helping those in corporate finance understand the valuation of their own firm. Stock-price data are also important for evaluating potential investments, which helps the analyst recommend buying, selling, or holding another company's stock. Current and historical daily stock-return data are available from multiple sources, including **Yahoo! Finance**, **MarketWatch**, **Bloomberg**, **NASDAQ**, and **NYSE**. Exhibit 9.2 provides an example of historical stock-price data for **Chipotle** available in MarketWatch.



DATE	OPEN	HIGH	LOW	CLOSE	VOLUME
11/30/2021	$1,670.32	$1,682.11	$1,635.00	$1,643.41	270,574
11/29/2021	$1,702.45	$1,702.45	$1,672.17	$1,678.10	141,529
11/26/2021	$1,678.15	$1,703.10	$1,671.12	$1,686.88	141,657
11/24/2021	$1,690.55	$1,716.67	$1,680.00	$1,707.93	108,965
11/23/2021	$1,715.30	$1,730.00	$1,671.50	$1,708.47	374,454
11/22/2021	$1,766.11	$1,766.99	$1,704.86	$1,721.42	240,347
11/19/2021	$1,780.05	$1,780.05	$1,752.75	$1,754.90	250,081

Exhibit 9.2
MarketWatch as Source of Current and Historical Daily Stock Price Data for Chipotle (Accessed December 1, 2021)

MarketWatch, Inc

Summarized Financial Data

Bloomberg, **Reuters**, **FactSet**, and **Quandl** all provide financial summarized data for a fee. However, some summarized financial data are available for free, such as a data source maintained by Aswath Damodaran, a New York University finance professor. Damodaran maintains up-to-date summary data of stock returns, risk measures, commodities prices, insider and institutional holdings, debt ratios, profit margins, and other financial data by industry on his website, http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/margin.html. This website and other similar finance data websites are used extensively in the finance industry to measure companies' past financial performance and predict their future financial performance.

Financial Statement Data

Much of the data needed to perform financial analytics is similar to the financial-statement data also used in accounting analytics (for a review, see Chapter 8). Also helpful are the repositories for financial data held at the Securities and Exchange Commission, including EDGAR and XBRL.

EDGAR: The Electronic Data Gathering, Analysis, and Retrieval System

The 10-K, 10-Q, 8-K, and other required submissions to the Securities and Exchange Commission (for companies publicly listed on stock exchanges; see Chapter 8) are all stored on **EDGAR** (https://www.sec.gov/edgar.shtml), the SEC's Electronic Data Gathering, Analysis, and Retrieval system. The EDGAR system automatically collects, validates, indexes, and serves as a repository for all forms submitted to the SEC. Exhibit 9.3 is an example of the items included in the EDGAR repository for **DoorDash**.



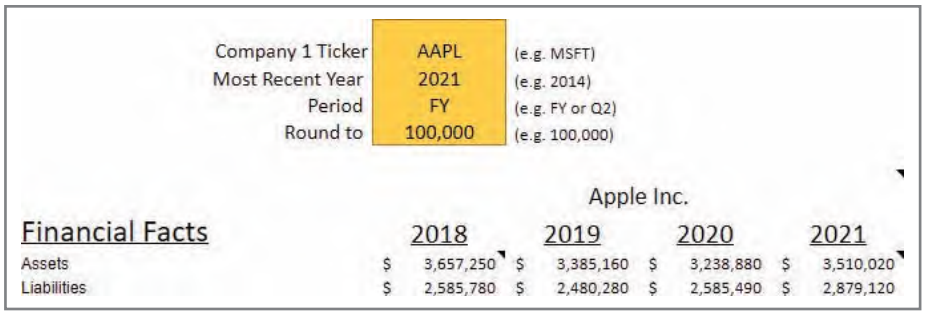
Exhibit 9.3
Example of the SEC Filings for DoorDash (as shown on EDGAR)

(Source: U.S Securities and Exchange Commission, EDGAR, https://www.sec.gov/cgi-bin/browse-edgar?company=doordash&match=&filenum=&State=&Country=&SIC=&myowner=exclude&action=getcompany.)

XBRL (eXtensible Business Reporting Language)

XBRL, which stands for **eXtensible Business Reporting Language,** is the computer-based standard used to define and exchange financial information between disclosing companies (such as **Airbnb** and **Rivian**) and financial statement users (such as investors, regulators, lenders, banks, and bondholders) and other interested parties. XBRL requires most numbers from a financial statement to be accompanied by a tag from the XBRL library that classifies the reported number into a predefined, standardized schema.

The SEC requires publicly traded companies to submit their financial statements using XBRL. This consistency provides the SEC with a huge data set of financial-statement numbers that are available to interested parties. To access the XBRL data, the user inputs a company's ticker symbol, the relevant years or quarters, the time period (quarterly, entire year, or fiscal year), rounding preference (in $millions or $thousands), and the particular financial statement item as an XBRL tag. Exhibit 9.4 shows XBRL requesting data for **Apple**'s total assets (XBRL tag: Assets) and total liabilities (XBRL tag: Liabilities) from 2018 to 2021.



Company 1 Ticker	AAPL	(e.g. MSFT)
Most Recent Year	2021	(e.g. 2014)
Period	FY	(e.g. FY or Q2)
Round to	100,000	(e.g. 100,000)

Apple Inc.

Financial Facts	2018	2019	2020	2021
Assets	$ 3,657,250	$ 3,385,160	$ 3,238,880	$ 3,510,020
Liabilities	$ 2,585,780	$ 2,480,280	$ 2,585,490	$ 2,879,120

Exhibit 9.4
Use of XBRL Tags and the XBRL Data Retrieved (for Apple)

Inline XBRL/U.S. Securities and Exchange Commission

PROGRESS CHECK

7. What is XBRL? How are the XBRL data useful to a financial analyst?
8. What is the difference between the required filings held in the SEC's EDGAR database and XBRL?

9.4 DESCRIPTIVE FINANCIAL ANALYTICS

LO 9.4

Describe and provide examples of descriptive analytics used in financial analysis.

After the question is specified and the financial data are ready for analysis, the next step in the SOAR analytics model is to analyze the data. Recall that descriptive analytics characterizes, summarizes, and organizes features and properties of the data to facilitate understanding of financial performance results and the underlying data that support it.

In this section, we focus on descriptive financial analytics. Later sections focus on diagnostic, predictive, and prescriptive financial analytics.

Statistical and Summarization Techniques for Descriptive Analytics

Financial analysts use *summarization and statistical techniques* to address the basic questions "What happened?" and "What is happening?" Counts, totals, sums, averages, subtotals, minimums, maximums, histograms, pivot tables, standard deviations, and ratios are all descriptive analytics techniques available within Excel to characterize, summarize, and organize the data. Exhibit 9.5 provides a summary of these techniques.

Statistical Technique (Descriptive Analytics)	Description	Example Finance Question
Counts	Show how frequently an event occurs	How many companies have gone public (had an initial public offering) in the past six months?
Totals, sums, averages, subtotals	Summarize measures of performance	How many loans did a bank extend in the previous quarter?
Minimums, maximums, medians, standard deviations	Summarize measures showing extreme values to help explain what happened	Which stock in the portfolio has the highest price volatility (standard deviation)? Which stock in the portfolio has the lowest price volatility (standard deviation)?
Graphs (bar charts), histograms	Show performance graphically	What are the levels of quarterly profits over the past three years as shown in a bar graph?
Ratio analysis such as price-to-earnings, price-to-book, and price-to-sales ratios	Calculate important financial ratios for comparison between companies	What are the relative market valuations of electric vehicle manufacturers such as **Rivian**, **Canoo**, **Tesla**, and **Lucid**? Does a company have a higher price-to-earnings ratio than its competitors?

Exhibit 9.5
Statistical Techniques Used in Descriptive Financial Analytics to Address Finance Questions

Dividends: A sum of money paid regularly (typically quarterly) by a company to its shareholders out of its profits.

Baa corporate bonds: High-quality corporate bonds.

Descriptive Statistics

Descriptive statistics provide the basic statistics for financial figures. For example, Exhibit 9.6 shows the descriptive statistics for annual returns to the S&P 500 including dividends, Baa corporate bonds, and U.S. Treasury bonds since 1928, displaying the measures of central tendency (mean and median) and dispersion (standard deviation, variance, and range). The S&P 500 is composed of the largest 500 publicly traded firms. For a refresher on the statistics reported in Exhibit 9.6, see Chapter 3.

S&P 500 (includes dividends)		*Baa Corporate Bonds*		*US Treasury Bonds*	
Mean	0.116416129	Mean	0.072530108	Mean	0.052123656
Standard Error	0.020208469	Standard Error	0.007802306	Standard Error	0.007939111
Median	0.1422	Median	0.065	Median	0.0329
Mode	#N/A	Mode	#N/A	Mode	#N/A
Standard Deviation	0.194883416	Standard Deviation	0.075242719	Standard Deviation	0.076562019
Sample Variance	0.037979546	Sample Variance	0.005661467	Sample Variance	0.005861743
Kurtosis	0.119804299	Kurtosis	0.648403241	Kurtosis	1.671185568
Skewness	−0.426542856	Skewness	0.277235338	Skewness	0.980997076
Range	0.964	Range	0.4473	Range	0.4393
Minimum	−0.4384	Minimum	−0.1568	Minimum	−0.1112
Maximum	0.5256	Maximum	0.2905	Maximum	0.3281
Sum	10.8267	Sum	6.7453	Sum	4.8475
Count	93	Count	93	Count	93

Exhibit 9.6
Descriptive Statistics for Annual Returns to S&P 500 (including dividends), Baa Corporate Bonds, and U.S. Treasury Bonds Since 1928

(Source for Data: http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/histretSP.html. Accessed 4/23/2020. Source for Analysis: Excel Data Analysis ToolPak: Descriptive Statistics.)

LAB CONNECTION

LAB 9.1 EXCEL asks you to use Excel's Data Analysis ToolPak to create the descriptive statistics for annual returns (including dividends) to the S&P 500 stock index, Baa corporate bonds, and U.S. Treasury bonds since 1928.

PROGRESS CHECK

9. What are the benefits of using a bar graph to show a company's stock-market returns over the past five years?
10. What descriptive statistics might be used to describe a bank's total outstanding loans? Beyond just the total amount of loans outstanding, what other information would be helpful to describe the loans outstanding?

9.5 DIAGNOSTIC FINANCIAL ANALYTICS

LO 9.5

Describe the types of finance questions addressed by diagnostic analytics.

Diagnostic analytics attempts to answer the questions "Why did it happen?" and "What are the causes of past results?" Recall that diagnostic analytics investigates the underlying cause that is not apparent in the descriptive data. Diagnostic analytics uses various types of analyses, including finding anomalies and outliers and using drill-down analytics to determine relations, patterns, and/or linkages among variables. In Exhibit 9.7 and the subsequent sections, we provide examples of how each method can be used in financial analytics.

Categories of Diagnostic Analytics	Example Finance Question	Specific Technique Employed
Finding anomalies/ outliers	How has **General Motors** stock performed relative to other auto makers and the economy?	Evaluate stock market performance relative to benchmarks.
Drill-down analytics: DuPont analysis	How can we explain recent performance at **Disney**? Was it due to increased profit margin, efficiency, or financial leverage, or a combination of these factors?	Perform DuPont analysis to explain return on equity (ROE) or return on assets (ROA).
Drill-down analytics: Sharpe ratio	To what extent is the return on investment related to increased risk?	Use the Sharpe ratio to evaluate the relationship between risk and return.

Exhibit 9.7
Categories of Diagnostic Analytics, Example Finance Questions, and Specific Techniques Employed

Finding Anomalies and Outliers

Diagnostic analytics is sometimes used to explain why some phenomenon is different than expected. Before we can explain why something is different, we have to identify those items that differ from expectations or from a benchmark. **Anomalies** are observations or data points that deviate from the expected or norm, and **outliers** are observations or data points that lie outside their expected distribution.

Exhibit 9.8 provides an example of diagnostic analytics by providing a line graph of performance compared to an expectation or benchmark. The exhibit reveals that **Walmart** did not perform as well as the S&P 500 or its retailing benchmark over the years 2017–2022. A $100 investment in Walmart stock made in 2017 would have been worth $231.82 by January 31, 2022, but a $100 investment in the benchmark S&P 500 Retailing Index would have been worth $296.49 in early 2022. The question is: What are the reasons for Walmart's performance? Knowing that Walmart underperformed its industry and overperformed the stock market overall encourages the analyst to look for explanations. Is the performance due to the e-commerce strength of Amazon or the pandemic? Is it due to Walmart leaving Brazil or slowing down its sales in China? Answering all of these questions requires diagnostic analytics.



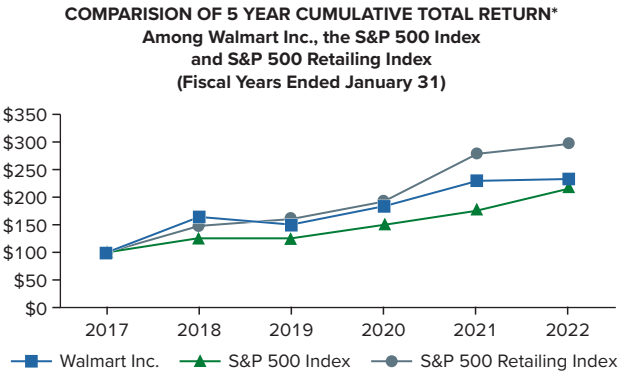
DATA VISUALIZATION

Drill-Down Analytics Using DuPont Ratios

Ratio analysis is a quantitative analysis of financial-statement items to help evaluate a company's liquidity, solvency, profitability, and efficiency. Ratio analysis is sometimes used as a descriptive analytics technique, but it can also be used as a diagnostic analytics technique to drill down further to determine the underlying reasons for a company's performance.

Stock Performance Chart

This graph compares the cumulative total shareholder return on Walmart's common stock during the five fiscal years ended through fiscal 2022 to the cumulative total returns on the S&P 500 Retailing Index and the S&P 500 Index. The comparison assumes $100 was invested on February 1, 2017 in shares of our common stock and in each of the indices shown and assumes that all of the dividends were reinvested.



*Assumes $100 Invested on Februrary 1, 2017
Assumes Dividends Reinvested
Fiscal Year ended January 31, 2022

	Fiscal Years Ended January 31					
	2017	2018	2019	2020	2021	2022
Walmart Inc.	$ 100.00	$ 163.83	$ 150.78	$ 183.78	$ 229.31	$ 231.82
S&P 500 Index	100.00	126.41	123.48	150.26	176.18	217.21
S&P 500 Retailing Index	100.00	148.34	159.89	190.43	278.09	296.49

Exhibit 9.8
Returns to Investment in Walmart Stock Relative to Benchmarks

(Source: Walmart 2020 Annual Report.)

One method of drill-down analytics used in finance uses DuPont ratio analysis to better understand the return on stockholders' **equity**, which is defined as the value of shareholder investments in the company and calculated as total assets minus total liabilities. **DuPont ratio analysis**, developed at and for the DuPont Corporation, summarizes and disaggregates company performance into three ratios to help explain the return on stockholders' equity and address the important question "What happened?" These three ratios are:

1. **Profit margin** or *profitability,* calculated as net income divided by total sales revenues. This ratio identifies the extent to which the company earns income on each sales dollar.
2. **Asset turnover**, calculated as sales divided by total assets. This ratio is a measure of activity or efficiency.
3. **Financial leverage**, calculated as total assets divided by stockholders' equity.[2] This ratio represents the extent to which assets are purchased using debt.

[2]Average total assets are often used instead of total assets at the end of the year. We use total assets at the end of the year for our examples here and in this chapter's labs. In addition, total stockholders' equity at the end of the year is used instead of average stockholders' equity for our examples here and in the labs.

Return on equity (ROE) is a measure of profitability generated for each dollar of stockholders' investment. Generally, it is defined as net profit (income) divided by stockholders' equity, or Net profit (income) ÷ Equity. In DuPont ratio analysis, ROE is calculated as follows[2]:

ROE
= Profit margin × Asset turnover × Financial leverage
= (Net profit/Sales) × (Sales/Total assets) × (Total assets/Stockholders' equity)

Alphabet (in $ millions)				
Year	**Sales Revenue**	**Net Profit (Net Income)**	**Assets**	**Stockholders' Equity**
2020	$182,527	$40,269	$319,616	$222,544
2019	$161,857	$34,913	$275,909	$201,442
2018	$136,819	$30,736	$232,792	$177,628
2017	$110,855	$12,662	$197,295	$152,502
2016	$90,272	$19,478	$167,497	$139,036
2015	$74,989	$16,348	$147,461	$120,331
2014	$66,001	$14,136	$129,187	$103,860
2013	$55,519	$12,733	$110,920	$86,977
2012	$46,039	$10,737	$93,798	$71,570

Exhibit 9.9
Data for Analysis of Alphabet (Google's Parent Company) Financial Statement Data Using Calcbench (XBRL) Data

(Source: Alphabet Financial Statements 2012-2020.)

Exhibit 9.9 summarizes key information from **Alphabet**'s (Google's parent company) financial statements. Let's use these numbers to calculate each of the DuPont ratios.

The 2020 profit margin is computed as Net profit/Sales = $40,269/$182,527 = 0.2206.

The 2020 asset turnover is computed as Sales/Assets = $182,527/$319,616 = 0.5711.

The 2020 financial leverage is computed as Assets/Equity = $319,616/$222,544 = 1.4362.

The 2020 return on equity (ROE) is computed as Net profit/Equity = $40,269/$222,544 = 0.1809, or 18.09%.

Using the DuPont ratio, note that in 2020 Profit Margin (0.2206) * Asset turnover (0.5711) * Financial leverage (1.4362) = Return on equity (0.1809).

Exhibit 9.10 provides a summary of the DuPont analysis. Use the formulas above to check the numbers for 2018 and 2019.

Abbreviated DuPont Ratio Analysis for Return on Assets

Sometimes DuPont analysis is used to determine the return on assets instead of return on equity. The analysis is similar, but it uses only the first two ratios to calculate the return on assets. **Return on assets (ROA)** is a measure of profitability for each dollar of resources (assets) available for management to use, computed as net profit (net income) ÷ total assets.

Exhibit 9.10
Solutions for DuPont Ratios Analysis of Google (Alphabet) Using Calcbench (XBRL)

(Data Source: https://www.calcbench.com/xbrl_to_excel.)

Google (Alphabet) ($ millions)				
Year	**Profit Margin**	***Asset Turnover**	***Financial Leverage**	**= Return on Equity**
2020	0.2206	0.5711	1.4362	0.1809
2019	0.2157	0.5866	1.3697	0.1733
2018	0.2246	0.5877	1.3106	0.1730
2017	0.1142	0.5619	1.2937	0.0830
2016	0.2158	0.5389	1.2047	0.1401
2015	0.2180	0.5085	1.2255	0.1359
2014	0.2142	0.5109	1.2439	0.1361
2013	0.2293	0.5005	1.2753	0.1464
2012	0.2332	0.4908	1.3106	0.1500

Here, we will compare **Walmart** to **Target**, two retail competitors, over a recent year. The DuPont analysis for return on assets is:

Return on assets	= Profit margin	× Asset turnover
	= Net profit (Net income)/Sales	× (Sales/Total assets)
Walmart's return on assets (2020)	= Profit margin	× Asset turnover
	= ($13,510/$559,151)	× ($559,151/$252,496)
	= 2.42%	× 2.2145
	= 5.35%	
Target's return on assets (2020)	= Profit margin	× Asset turnover
	= ($4,368/$93,561)	× ($93,561/$54,248)
	= 4.67%	× 1.7247
	= 8.05%	

Whereas profit margin tells you how much profit a company makes per dollar of sales, asset turnover tells you how efficiently a company uses its assets. In our example, Walmart has a lower profit margin (2.42%) on each dollar of sales compared to Target (4.67%). However, Walmart has a faster asset turnover (2.2145) than Target (1.7247). For this year of comparison, Target's ROA (8.05%) is higher than Walmart's (5.35%).

Not only will you see dramatic differences in profit margins and asset turnover within an industry, you'll also see differences among industries. For example, toilet paper has a relatively small profit margin but a high asset turnover. In contrast, jewelry has a relatively high profit margin but a much smaller asset turnover. These varying industry profit margins and asset turnover ratios help serve as an expectation to be used as a threshold for evaluating company performance.

As an illustration of the varying industry profit margins, Aswath Damodaran maintains an up-to-date listing of profit margins by industry at http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/margin.html.

LAB CONNECTION

LAB 9.2 EXCEL asks you to conduct a DuPont analysis for **Smucker's**, **Hershey**, **Kellogg**, and **Kraft Heinz**. It gives the relevant underlying financial-statement figures and then asks you to compute return on equity and return on assets.

Risk/Return and the Sharpe Ratio

There is a presumed relationship between risk and expected return on an investment: The higher the risk, the higher the amount of expected return (Exhibit 9.11). For example, U.S. Treasury bonds and U.S. Treasury bills have no risk, because the government has the authority to print more money to pay its bills. A certificate of deposit (CD) at a bank has very low risk because it is generally insured by the government. Corporate bonds are considered to have mid-level risk, and thus investors expect a medium level of return to an investment in corporate bonds. Investments in a stock index fund (such as a Vanguard market index) or an individual stock represent much higher levels of risk and thus a much higher expected return.

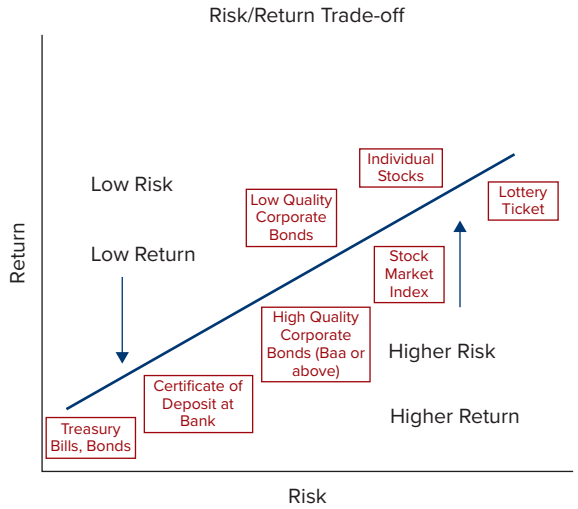


Exhibit 9.11
The Risk/Return Trade-Off

Although these relationships between risk and return exist in theory, we can use two different drill-down techniques to assess whether this relationship exists in practice: the Sharpe ratio and regression analysis.

SOARing to Success

Assessing the Risk of All Possible Investments

Because of the relationship between risk and return, it is important for finance professionals to assess the risk of each possible investment. Specifically, finance professionals assess the risk of not seeing either a return on the investment made or a return of the original investment made.

Risk/Return Diagnostic Analysis Using the Sharpe Ratio

One way to evaluate risk and return is to use the Sharpe ratio. The **Sharpe ratio**, developed by Nobel laureate William F. Sharpe, is a ratio of the average return earned in excess of the risk-free rate (on 10-year Treasury bonds) to a unit of volatility or total risk. It is widely used to evaluate relative risk and return.

Exhibit 9.12 shows that the average annual return is greater for the S&P 500 than the return on Baa corporate bonds. It also shows that one measure of risk (standard deviation of annual returns) is greater for the S&P 500 Stock Market Index than it is for the Baa corporate bonds. These numbers suggest that risk is greater for stock than for Baa-rated corporate bonds. However, we do not know if the compensation (excess rate of return) per unit of risk is greater for either of them. To find that information, we compute the Sharpe ratio as follows for the S&P 500 Stock Market Index and Baa-rated corporate bonds:

$$\text{Sharpe Ratio (S\&P500)} = \frac{\text{Average Annual Return for S\&P 500 Index} - \text{Risk Free Rate (Treasury Bonds)}}{\text{Standard Deviation of Annual Returns on S\&P 500}}$$

$$\text{Sharpe Ratio (Baa Corporate Bonds)} = \frac{\text{Average Annual Return for Baa Corporate Bonds} - \text{Risk Free Rate (Treasury Bonds)}}{\text{Standard Deviation of Annual Returns on Baa Corporate Bonds}}$$

Exhibit 9.12
Sharpe Ratio for Both the S&P 500 Stock Market Index and Baa Corporate Bonds (as computed in Lab 9.1 Excel)

	D	E	F	G
97				
98			S&P 500	Baa Corporate Bonds
99	Avg. (Excess) Return		6.71%	2.08%
100	Standard Deviation		21.09%	6.83%
101	Sharpe Ratio		31.82%	30.40%

As the "Sharpe Ratio" row in Exhibit 9.12 shows, the Sharpe ratio for the S&P 500 and Baa corporate bonds suggests a very close relationship between risk and return for both of them, suggesting that the risk is compensated in a similar way by the returns to both the S&P 500 and the Baa corporate bonds. It also provides some evidence that there is a relationship between risk and return—specifically, the higher the risk, the higher the return (as measured by standard deviation of the returns) needed to compensate investors for taking that risk.

LAB CONNECTION

LAB 9.1 EXCEL asks you to calculate the Sharpe ratio for two assets that have traded for the past 90-plus years: the S&P 500 and Baa-rated corporate bonds.

Risk/Return Diagnostic Analysis Using Regression Analysis

Recall from earlier chapters that we can use regression analysis to determine the relationship between two variables when we believe one variable is dependent on other

independent variables. In addition to the Sharpe ratio, we can use regression to analyze the risk/return relationship, as follows:

- Dependent (outcome) variable: Different measures of expected investment return, such as past stock prices or returns, bond yields, or returns on real estate investments
- Independent variables: Different measures of risk, such as past stock price volatility, volatility of earnings, bankruptcy risk, and level of debt

Once we learn the past and current relationship between risk and return, we can use predictive analytics to predict future stock returns based on risk. In finance, the **capital asset pricing model (CAPM)** is used to predict the relationship between systematic risk and expected return for financial assets. Explained in more detail in advanced finance courses, CAPM is most often used for stock price valuation. Investors can also use it to evaluate the risk and returns of other assets, including bonds, commodities, and real estate.

PROGRESS CHECK

11. What are the three components of the DuPont analysis, and how does it help explain the return on equity? Why is DuPont analysis considered to be a diagnostic analysis technique?
12. How is the Sharpe ratio useful for evaluating the relative risk and return of various investments?

9.6 PREDICTIVE FINANCIAL ANALYTICS

LO 9.6

Describe the types of finance questions addressed by predictive analytics.

Predictive analytics forecasts or predicts future events or performance. Exhibit 9.13 summarizes three predictive analytics techniques, along with example questions and the technique used. In the following sections, we show how classification and regression can be applied to financial analytics. (We discussed forecasting with time series analysis in Chapter 8; see Learning Objective 8.6.)

Predictive Analytics Technique	Example Finance Question	Specific Technique Employed
Classification: Separates or classifies a sample (or population) into two or more groups or classes	Which potential borrowers will most likely pay back their loan (and which will not)? Which customers should be granted a bank loan (and which should not)?	Run classification analysis to predict which borrowers will pay back a loan and which will not.
Regression: Predicts a specific output value based on variable inputs	Can we predict the interest rate given to prospective borrowers at a bank?	Run regression analysis to predict the interest rate based on the prospective borrower's characteristics.
Time series analysis: Predicts future values based on past values of the same variable	Can we predict next year's dividends?	Use time series analysis incorporating past values of dividends to predict the next few years' dividends.

Exhibit 9.13
Predictive Analytics Techniques, Finance Examples, and Specific Techniques Employed

Altman's Z and Bankruptcy Classification

In 1968, Edward Altman published a paper classifying a company's bankruptcy risk using an analysis now called *Altman's Z*.[3] The basic calculation is still used today, more than 50 years later. This classification exercise is so useful because it allows analysts to put companies into classes according to their risk of bankruptcy. Investors are particularly interested in this information, as are banks and auditors. For example, investors are unlikely to purchase a company's stock if there is a strong chance it will go bankrupt. Auditors will not give a clean audit opinion of a company that is about to go bankrupt, and banks will not want to extend loans to companies that are teetering on the edge of bankruptcy.

The foundation of Altman's bankruptcy analysis uses common business ratios to compare companies with a high risk of bankruptcy to a matched sample of non-bankrupt companies. Specifically, Altman found that a combination of five common business ratios, summarized in Exhibit 9.14, help to predict bankruptcy. The higher each of these common business ratios (X_1:X_5), the less likely the company will go bankrupt.

Factor Number	Business Ratio Predicting Bankruptcy	Measures
X_1: Liquidity	Working capital (very short-term assets similar to cash) ÷ Total assets	The company's ability to pay its current bills
X_2: Long-term profitability	Retained earnings (accumulated earnings not yet paid to shareholders) ÷ Total assets	The company's profitability over the long run
X_3: Short-term profitability	Earnings before interest and taxes ÷ Total assets	Whether the company is currently profitable
X_4: Long-term solvency	Market value of stock ÷ Book value of liabilities	The company's ability to pay its debt as it comes due
X_5: Asset efficiency	Sales ÷ Total assets	How well assets are utilized

Exhibit 9.14
Altman's Z Bankruptcy Factors

Altman performed his analysis based on 33 bankrupt manufacturing companies and 33 non-bankrupt manufacturing companies. His statistical analysis resulted in this formula for calculating a *z*-score (similar to the *z*-scores discussed in Chapter 3):

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$$

Altman found that Z for the bankrupt group averaged −0.25, while Z for the non-bankrupt group averaged 4.48. He then used the following decision rules to establish expectations for bankruptcy:

If $Z < 1.80$	⇒	Classify as significant risk of bankruptcy, the "distress zone"
If $Z \geq 1.80$ and $Z < 3.0$	⇒	Classify as at risk of bankruptcy, the "gray zone"
If $Z \geq 3.00$	⇒	Classify as not currently at risk of bankruptcy, the "safe zone"

We can apply Altman's Z-score and his decision rules to any company to evaluate its risk of bankruptcy. Let's consider the case of **Apple** Computer in 1997. At that time, Windows had just come out for the IBM PC, so many were questioning the value of Apple computers, which at that time did not interface well with PCs and were not commonly used in the corporate world. At that point, many believed Apple was on the verge of bankruptcy. Using Altman's Z-score, we can compute each of the five predictive factors of bankruptcy for Apple in 1997, multiplying them by their respective weights as noted in the Z-score formula ($Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$), as Exhibit 9.15 shows.

[3]Altman, Edward I., "Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy," *The Journal of Finance,* 23.4 (1968), 589-609.

Apple Computer	Year 1997	Altman Z Weights
X_1: Working capital (Current assets – Current liabilities) ÷ Total assets	**0.379**	*** 1.2 = 0.4548**
X_2: Retained earnings (Accumulated earnings not paid to shareholders) ÷ Total assets	**0.139**	*** 1.4 = 0.1946**
X_3: Earnings before interest and taxes ÷ Total assets	**−0.253**	*** 3.3 = −0.8349**
X_4: Market value of stock ÷ Book value of liabilities	**0.896**	*** 0.6 = 0.5376**
X_5: Sales ÷ Total assets	**1.673**	*** 1.0 = 1.6730**
	Altman Z-score: 2.0251	

Exhibit 9.15
Altman's Z-Score Formula Applied to Apple Inc. in 1997

Adding all the components together results in an Altman Z-score of 2.0251. The analysis shows that Apple was in the gray zone (where $Z \geq 1.80$ and $Z < 3.0$), very close to the 1.80 cutoff for bankruptcy. However, Steve Jobs rejoined Apple in July of that year, driving the development of iPods, and later, iPhones and iPads, which turned Apple into one of the world's most valuable companies.

LAB CONNECTION

LAB 9.4 EXCEL provides step-by-step discussion, analysis, and practice with classifying companies as at significant risk of bankruptcy, at risk of bankruptcy, or at low risk of bankruptcy.

This type of bankruptcy analysis can be applied to companies and industries facing dramatic changes due to changes in technology or competition. For example, in recent years, the retail industry has been undergoing a transformation primarily because **Amazon**'s e-commerce initiatives have placed pressure on many traditional bricks-and-mortar retail companies to change their business model or go bankrupt. For example, an analysis for 2018 of 5,437 companies in the retail industry suggests the bankruptcy risk shown in Exhibit 9.16. While more than half of retail firms were not at risk of bankruptcy, about 35 percent of retail firms were at significant risk of bankruptcy during this period of significant change in retail.

Bankruptcy Zone Based on Altman's Z	Number of Firms
At significant risk of bankruptcy (distress zone)	2,530
At risk of bankruptcy (gray zone)	817
Not at risk of bankruptcy (safe zone)	2,090
Total companies included in analysis	5,437

Exhibit 9.16
Retail-Industry Analysis of Firms, 2017–2018: Bankruptcy Zone Based on Altman's Z

Predicting Who Will Be Offered a Loan



An important decision for lenders is whether to extend a loan to a potential borrower. Banks and other lenders evaluate borrowers based on their creditworthiness to decide whether they will extend a loan to those borrowers. First and foremost, the banks are trying to determine if the potential borrower will be able to repay the loan. For example, to help predict whether individuals will repay their loans, lenders use various borrower characteristics, including credit scores, work history, existing levels of total debt, debt-to-income ratios, and whether they own or rent their home. Although these characteristics are commonly used, it should be noted that some critics have questioned the ethics of making

Exhibit 9.17
Lending Club: A Peer-to-Peer Lender

LendingClub Bank



decisions based on them, as they may unfairly discriminate against people from lower socioeconomic classes.

As an example of the analytics involved in deciding whether to extend a loan, consider the case of **Lending Club**, a U.S.-based peer-to-peer lending company, headquartered in San Francisco, California (Exhibit 9.17). Lending Club uses its past loan data to classify its borrowers into two classes: those who should be given a loan and those who should not, based on the borrower's likelihood of paying back the loan made by the Lending Club. The following are some independent variables that might be used to classify Lending Club's potential borrowers:

Credit History

- Does the borrower have any delinquent accounts (where the payments are late or behind schedule)?
- What is the borrower's monthly credit-card bill?
- Does the borrower have any unpaid collections accounts, such as unpaid medical bills?
- Has the borrower declared bankruptcy in the past three to five years?
- How many recent credit applications has the potential borrower made?
- How much total outstanding debt does the borrower have?
- What is the borrower's debt-to-annual income ratio?
- What is the borrower's **credit score** (a statistic that evaluates a customer's creditworthiness and ability to repay a loan)?
- Does the borrower rent or own their residence?

Income

- What is the borrower's annual income? Is the income verified?
- How long has the borrower been employed at their current job?
- Does the borrower's income come from self-employment, or does it come from salary?
- What is the borrower's salary history?

Loan Specifics

- For what purpose is the borrower intending to use the loan?
- How much money is the borrower requesting?

After considering these questions, the analyst might believe that three variables will most likely predict loan acceptance or loan rejection: amount of requested loan, employment length, and debt-to-annual income ratio. Assuming these are the variables chosen, what is the possible relationship between these variables and loan acceptance or rejection?

1. *Amount of the requested loan (loan_amnt).* As the requested loan amount becomes larger, do you think the chances of Lending Club granting the loan will be higher or lower? In general, the bigger the loan, the less likely the loan will be given, so we expect a negative (−) relationship between loan granting and loan amount.
2. *Employment length (Emp Length).* As the prospective borrower's employment length becomes longer, do you think the chances of Lending Club granting the loan will be higher or lower? In general, the longer the employment length, the more likely the loan will be given, due to the stability of the borrower's job. Thus, we would expect a positive (+) relationship between loan granting and employment length.
3. *Debt-to-annual income ratio (DTI Bucket).* As the debt-to-annual income ratio of the prospective borrower becomes higher, are the chances of Lending Club granting a loan higher or lower? In general, the bigger the debt relative to income, the less likely the loan will be granted. Thus, we would expect a negative (−) relationship between loan granting and debt-to-annual income ratio. At Lending Club, the level of debt-to-annual income is summarized in high, medium, and low debt-to-annual income buckets.

Exhibit 9.18 shows the results of this regression analysis (as performed in Lab 9.3 Excel), which confirms our expectations. The negative coefficient on the loan amount (loan_amnt) of -9.13E-07 (see dot #1 in Exhibit 9.18) suggests a negative relationship between the size of the loan amount and loan acceptance. The negative coefficient on debt as compared to income (DTI bucket) of -0.02278 (see dot #2 in Exhibit 9.18) suggests a negative relationship between the amount of debt and loan acceptance. The positive coefficient on employment length (Emp Length) of 0.024 (see dot #3 in Exhibit 9.18) suggests a positive relationship between the length of employment and loan acceptance.

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.303754968				
R Square	0.092267081				
Adjusted R Square	0.0922631				
Standard Error	0.220123451				
Observations	684122				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	3369.398179	1123.133	23179.2	0
Residual	684118	33148.48184	0.048454		
Total	684121	36517.88002			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	0.07561005	0.000815409	92.72652	0	0.074011875
❶ loan_amnt	-9.13573E-07	2.56369E-08	-35.635	7.3E-278	-9.6382E-07
❸ Emp Length	0.023968143	9.69916E-05	247.1157	0	0.023778042
❷ DTI Bucket	-0.022777529	0.000321645	-70.8158	0	-0.023407942

Exhibit 9.18
Regression Results of Loans Accepted by Lending Club

LendingClub Bank

LAB CONNECTION

LAB 9.3 EXCEL provides step-by-step discussion, analysis, and practice with classifying the prospective borrowers who will receive loans from Lending Club and those who will not. The lab asks you to determine whether the three factors discussed above are indeed helpful in separating the loan applicants into two classes: approved for a loan or rejected for a loan.

PROGRESS CHECK

13. Why is classifying the risk of bankruptcy considered to be predictive analytics?
14. Why is the size of the requested loan amount from a prospective borrower likely to be negatively associated with that borrower receiving a loan?

LO 9.7

Describe the basics of prescriptive financial analytics.

9.7 PRESCRIPTIVE FINANCIAL ANALYTICS

Prescriptive analytics addresses the questions "What should we do, based on what we expect will happen?" and "How do we optimize our performance based on potential constraints?" As the second question makes clear, *prescriptive analytics* is performed to identify the best possible options given constraints or changing conditions.

We use five prescriptive analytics techniques to help determine what we should do based on what we expect will happen:

1. What-if analysis, including marginal/incremental analysis
2. Cash flow analysis
3. Scenario analysis
4. Goal-seek analysis
5. Sensitivity analysis

Exhibit 9.19 summarizes these techniques, along with example questions and the specific techniques used.

Prescriptive Analytics Technique	Example Finance Question	Specific Technique Employed
Marginal/incremental analysis: What-if analysis that determines the change in profit associated with the cost or benefit of the next, or marginal, unit	Should **Moderna** manufacture or outsource the manufacture of its vaccines?	Estimate the financial implications of two scenarios: make vaccines or buy vaccines from another manufacturer.
Cash flow analysis: Evaluates future cash flows for potential investments or expenditures, typically using net present value (NPV) or internal rate of return (IRR)	What are the expected returns to investing in a corporate bond or stock?	Estimate the amounts, timing, and risk associated with the predicted cash flows of a new financial investment.
Scenario analysis: What-if analysis that predicts how changing multiple inputs at the same time will affect an outcome variable	How do we respond to various scenarios of a strike being threatened by our employee union?	Analyze all the potential scenarios to determine the overall impact of a strike, whether there is no strike, a minimal strike, or a strike by all union employees.
Goal-seek analysis: What-if analysis that determines the required input value(s) to get a desired output (that is, backwards analysis)	What returns are necessary for **Toyota** to break even on its newest hybrid factory?	Perform break-even analysis to determine the needed level of returns to break even on the hybrid factory.
Sensitivity analysis: What-if analysis that predicts how changing one input will affect an outcome variable	What happens to profits at **Bank of America** if interest rates change in the future from 3 percent to 4 percent or 5 percent or more?	Test to see how sensitive **Bank of America**'s profits are to interest-rate changes based on our predictive models.

Exhibit 9.19
Prescriptive Analytics Techniques, Example Finance Questions, and Specific Techniques Employed

Prescriptive Cash Flow Analysis

Using cash flow analysis, analysts forecast the amounts and timing of future cash flows. In other words, they predict both the size of future cash flows and when they expect the cash to arrive. There is also some uncertainty regarding when those future cash flows will arrive, and they factor this uncertainty into their analysis. Consistent with the risk and return discussed earlier in this chapter, the more uncertain (or risky) the cash flows, the higher the risk and the higher **discount** (or **interest) rate** used to evaluate the cash flows.

We can use two common methods to evaluate cash flows:

1. *Net present value (NPV).* By calculating the NPV using different rates of risk, we can determine where it makes sense to invest, given different levels of expected risk.
2. *Internal rate of return (IRR).* The IRR is the interest rate at which the present value of the cash outflows is exactly equal to the present value of the cash inflows. If the IRR is greater than the cost of the funds the company borrows, the company will consider the proposed investment in a positive light.

Let's look more closely at these two methods.

Net present value (NPV) is calculated as the present value of cash inflows minus the present value of the cash outflows, or

$$\text{Present value} = CF_t/(1 + r)^t$$

where

CF_t = cash flows for period t
r = the discount rate or interest rate (typically the firm's **cost of capital**, or cost of borrowing money that incorporates the level of risk)

Assume you want to purchase a new piece of equipment that is expected to cost $200,000 at the time of purchase (Year 0). The new equipment will return $150,000 per year for three years. Assume that the cost of capital (the cost of borrowing money), or the discount rate, is 10 percent (= .10). The NPV is calculated as follows:

Year 0	Present value = $-\$200,000/(1.10^0)$	=	**−$200,000**
Year 1	Present value = $\$150,000/(1.10^1)$	=	$136,364
Year 2	Present value = $\$150,000/(1.10^2)$	=	$123,967
Year 3	Present value = $\$150,000/(1.10^3)$	=	$112,697
	Sum of present values, or NPV		$173,028

Note that a cash outflow—in this case, the $200,000 up-front payment (or investment)—is included in the NPV analysis as a negative number. Cash inflows are included in the analysis as positive numbers, as reflected in years 1–3.

The **internal rate of return (IRR)** is the discount rate that makes the project's net present value equal to zero, with the present value of cash outflows equal to the present value of the cash inflows. While there is no exact formula to compute the internal rate of return, financial calculators and Microsoft Excel use an iterative technique to experiment with different rates of return until it reaches an NPV of exactly zero. The IRR for the above investment, computed in Excel, is 40.7%.

Evaluating Future Cash Flows: Net Present Value and Installment Payments

Let's dig more deeply into net present value with an easy example. Suppose you decide you want to buy a car. The car costs $20,000 at a 5 percent annual interest rate, requiring

annual payments to be made over five years (that is, one payment per year for five years). The loan is for $20,000. You can use the =PMT() function in Excel to find out what the annual payments will be.

To compute the yearly payment, input

= PMT(rate,nper,pv)

where

rate = the annual interest rate (5 percent in this case)
nper = the number of payments (five annual/yearly payments in this case)
pv = present value (the size of the loan desired = $20,000)

Inputting =PMT(0.05,5,20000) in Excel gives an annual payment of $4,619.50. In Exhibit 9.20, it appears as a negative figure, representing a cash payment or cash outflow that will need to be made.

Exhibit 9.20
Determining the Annual Amount of the Car Loan Payment

	A	B
1	$20,000.00	Car Loan
2	5%	Interest Rate
3		
4	($4,619.50)	Year 1
5		Year 2
6		Year 3
7		Year 4
8		Year 5

Next, we need to compute the payment for each of the five years, with the same amount paid each year (Exhibit 9.21).

Exhibit 9.21
Determining the Amount of the Car Loan Payment Each Year Over the Five-Year Term

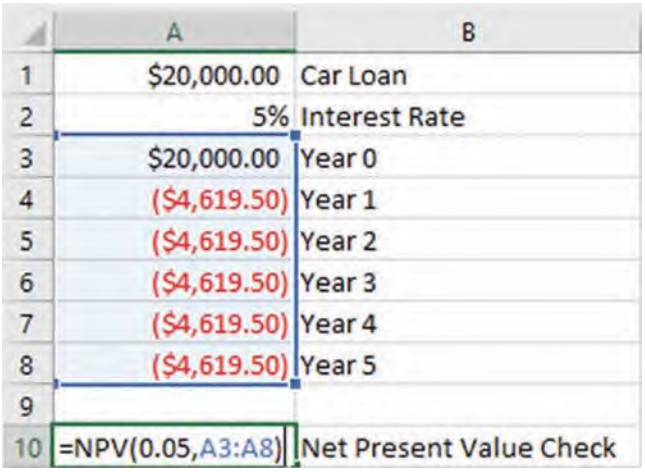
	A	B
1	$20,000.00	Car Loan
2	5%	Interest Rate
3		
4	($4,619.50)	Year 1
5	($4,619.50)	Year 2
6	($4,619.50)	Year 3
7	($4,619.50)	Year 4
8	($4,619.50)	Year 5

What is the net present value for these payments? In other words, what is the value of this string of five payments? It is $20,000. How do we verify that number?

First, input the $20,000 cash inflow we got to purchase the car as the Year 0 cash flow. Using the =NPV() function in Excel, compute the net present value of these cash flows. To do so, specify =NPV(rate, string of cash flows), where

rate = the annual interest rate (or discount rate)
string of cash flows = the amount of cash we expect to pay each year.

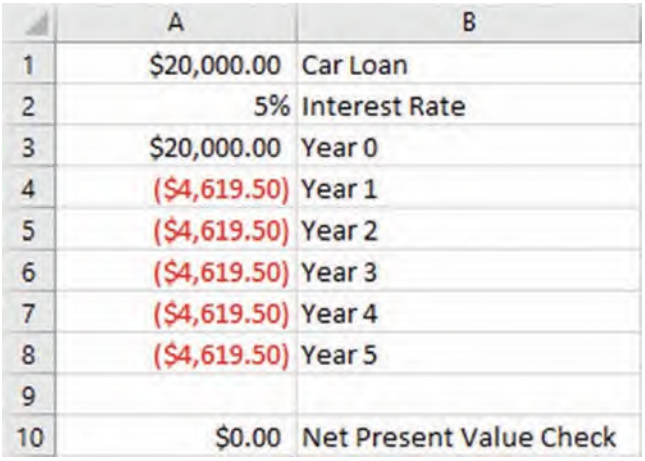
Exhibit 9.22 shows the calculation.



	A	B
1	$20,000.00	Car Loan
2	5%	Interest Rate
3	$20,000.00	Year 0
4	($4,619.50)	Year 1
5	($4,619.50)	Year 2
6	($4,619.50)	Year 3
7	($4,619.50)	Year 4
8	($4,619.50)	Year 5
9		
10	=NPV(0.05,A3:A8)	Net Present Value Check

Exhibit 9.22
Calculating the Net Present Value Over Years 0–5

As expected, the result will be an NPV of $0 (Exhibit 9.23).



	A	B
1	$20,000.00	Car Loan
2	5%	Interest Rate
3	$20,000.00	Year 0
4	($4,619.50)	Year 1
5	($4,619.50)	Year 2
6	($4,619.50)	Year 3
7	($4,619.50)	Year 4
8	($4,619.50)	Year 5
9		
10	$0.00	Net Present Value Check

Exhibit 9.23
Checking the Net Present Value Over Years 0–5

Recall that the internal rate of return is the rate that makes the net present value equal to $0 (as shown in Row 10 of Exhibit 9.24), which is computed to be 5 percent (as shown in cell A11 of Exhibit 9.25). Therefore, someone with a 5 percent interest rate would be indifferent between receiving $20,000 cash today (as a loan to pay for the car) or $4,619.50 each year for five years, resulting in a net present value of $0. At this **point of indifference**, it does not matter which alternative is chosen because the outcome is the same under either alternative.

Exhibit 9.24
Calculating the Internal Rate of Return

	A	B
1	$20,000.00	Car Loan
2	5%	Interest Rate
3	$20,000.00	Year 0
4	($4,619.50)	Year 1
5	($4,619.50)	Year 2
6	($4,619.50)	Year 3
7	($4,619.50)	Year 4
8	($4,619.50)	Year 5
9		
10	$0.00	Net Present Value Check
11	=IRR(A3:A8)	Internal Rate of Return

Exhibit 9.25
The Internal Rate of Return Over the Life of the Car Loan (Years 0–5)

	A	B
1	$20,000.00	Car Loan
2	5%	Interest Rate
3	$20,000.00	Year 0
4	($4,619.50)	Year 1
5	($4,619.50)	Year 2
6	($4,619.50)	Year 3
7	($4,619.50)	Year 4
8	($4,619.50)	Year 5
9		
10	$0.00	Net Present Value Check
11	5%	Internal Rate of Return

Evaluating Future Cash Flows: Capital Budgeting and Investment Decisions

Prescriptive analysis works to identify the best possible options given constraints. These constraints, such as a lack of unlimited funding for potential projects or investments, require companies to evaluate which projects might pay off at the highest level. **Capital budgeting** is the process of evaluating potential (typically large) investments to help identify which investment, if any, a company should choose.

Suppose that **Corey Stanford Consulting** is evaluating three possible investments in software. Each potential investment costs a different amount and has a different profit payoff structure over its expected four-year life. The analysis suggests the string of cash flows shown in Exhibit 9.26.

	A	B	C	D	E
1	Year		Investment 1	Investment 2	Investment 3
2	Year 0 (today)	Buy New Software	(8,000)	(10,000)	(12,000)
3	Year 1	Increased Profits	2,500	3,000	4,000
4	Year 2	Increased Profits	2,500	3,000	4,000
5	Year 3	Increased Profits	2,500	3,000	4,000
6	Year 4	Increased Profits	2,500	3,000	4,000

Exhibit 9.26
Evaluating the Cash Flows to Three Different Possible Capital Investments

Note that Year 0 is today, or the date of the investment. Years 1–4 are years when Corey Stanford Consulting expects to receive profits from the investment. Which investment should Corey Stanford choose?

Using the NPV and IRR techniques can help Corey make his decision. As part of the NPV analysis, Corey needs to know his cost of capital to evaluate the cash flows. The cost of capital is the cost of available funds to Corey, and it acts as a hurdle rate that the company must overcome to create value. Exhibit 9.27 uses a cost of capital of 10 percent (see cell F2), which suggests that the undiscounted cash flows and the net present value for Investment 3 are the highest, at $12,679.46 (see row 10). Lab 9.5 Excel provides additional details on a similar problem.

	A	B	C	D	E	F
1	Year		Investment 1	Investment 2	Investment 3	Cost of Capital
2	Year 0 (today)	Buy New Software	(8,000)	(10,000)	(12,000)	10%
3	Year 1	Increased Profits	2,500	3,000	4,000	
4	Year 2	Increased Profits	2,500	3,000	4,000	
5	Year 3	Increased Profits	2,500	3,000	4,000	
6	Year 4	Increased Profits	2,500	3,000	4,000	
7						
8	Net Cash Flows over Life (not discounted)		2,000	2,000	4,000	
9						
10	Present Value of Expected Future Cash Inflows		$7,924.66	$9,509.60	$12,679.46	
11						
12	Initial Outlay		(8,000)	(10,000)	(12,000)	
13	Net Present Value (NPV)		($75.34)	($490.40)	$679.46	

Exhibit 9.27
Evaluating Potential Investments Using Net Present Value

Exhibit 9.28 also suggests that the internal rate of return for Investment 3 is the highest, at 12.59 percent. Lab 9.6 Excel provides additional details on a similar problem.

	A	B	C	D	E
1	Year		Investment 1	Investment 2	Investment 3
2	Year 0 (today)	Buy New Software	(8,000)	(10,000)	(12,000)
3	Year 1	Increased Profits	2,500	3,000	4,000
4	Year 2	Increased Profits	2,500	3,000	4,000
5	Year 3	Increased Profits	2,500	3,000	4,000
6	Year 4	Increased Profits	2,500	3,000	4,000
7					
8	Net Cash Flows over Life (not discounted)		2,000	2,000	4,000
9					
10	Internal Rate of Return (IRR)		9.56%	7.71%	12.59%

Exhibit 9.28
Evaluating Potential Investments Using Internal Rate of Return

MINI CASE: Amazon Plans to Build a Second Headquarters Building

After a nationwide frenzy in which **Amazon** received 238 proposals from cities around the country, the company announced plans to build a second headquarters building (nicknamed HQ2) in Arlington, Virginia. Following the COVID-19 pandemic, and in an era when working from home has become more prevalent, some may question the benefits of a second headquarters space. How can Amazon quantify the benefits of this proposed $2.5 billion capital investment? How will it use capital budgeting to evaluate the investment?

Decision-Making with Data

Will the new building give Amazon access to new markets or new products that will create revenues? Will it create cost savings among employees by creating synergies? How can Amazon compare these expected quantitative benefits using discounted cash flow or other methods to prioritize the potential capital investments that it might be considering? Are there additional qualitative benefits to consider? Explain.

Source: https://www.architecturaldigest.com/story/amazon-unveils-nature-infused-hq2-design-includes-helix. Accessed October 2021.

LAB CONNECTIONS

LABS 9.5 EXCEL and **9.6 EXCEL** provide additional examples of evaluating potential investments and projects using NPV and IRR techniques.

Sensitivity Analysis

Financial analytical models are used to assess the impact of changes in inputs. This chapter's opening story about **Husqvarna** provides an example of how financial analysts might use sensitivity analysis to assess the impact of a change in a single input variable. Now consider another example of how sensitivity analysis might be used: If the interest rate for the funds that **Wells Fargo** borrows increases by 2 percent, what is the impact on Wells Fargo's profitability and stock price valuation? Based on how sensitive the stock price is to interest rates, analysts may view Wells Fargo as too risky for investment, because a 2 percent increase in interest rates might cause Wells Fargo to be unprofitable, dramatically lowering its future profits, dividends, and stock price.

MINI CASE: What Will It Take to Get a $25,000 Electric Car at Tesla?

Tesla has announced that it will reduce the cost of its battery cells and packs to meet the company's goal of selling an electric car for $25,000. The price per kilowatt hour (kWh) is often used to measure the capacity of a battery pack in electric vehicles. Those prices have fallen over the last decade from $1,100/kWh to $156/kWh.

Analyzing the cost of the battery packs of the future has a dramatic impact on Tesla and all other electric car manufacturers. Financial analysts at Tesla continuously run what-if analysis, including sensitivity analysis and scenario analysis, to provide input to Elon

Musk and other decision-makers on the level of production and profitability at Tesla, which to a large degree may be determined by changing battery costs.

Decision-Making with Data

Some questions that Tesla management might ask are: (1) At what battery price point will a $25,000 electric car at Tesla become realistic? (2) Are there other variables that might affect product pricing, such as pricing by a host of competitors offering electric cars? What other questions might Tesla's managers ask regarding the company's future profitability?

Source: https://www.theverge.com/2020/9/22/21450916/tesla-battery-pack-elon-musk-price-kilowatt-hour-ev-cost-tabless. Accessed 10/26/2021.

PROGRESS CHECK

15. Why is cash flow analysis considered a type of prescriptive analytics?
16. If the internal rate of return (IRR) of an investment is 5 percent and the cost of funds is 6 percent, would you go ahead with the investment, or would you choose to forego the investment?

9.8 REPORT THE RESULTS

LO 9.8

Identify methods of reporting financial analytics results.

As we saw in Chapter 6, there are many ways to report results. Tables and data visualizations are particularly useful in reporting the findings of financial analytics to decision makers.



DATA VISUALIZATION

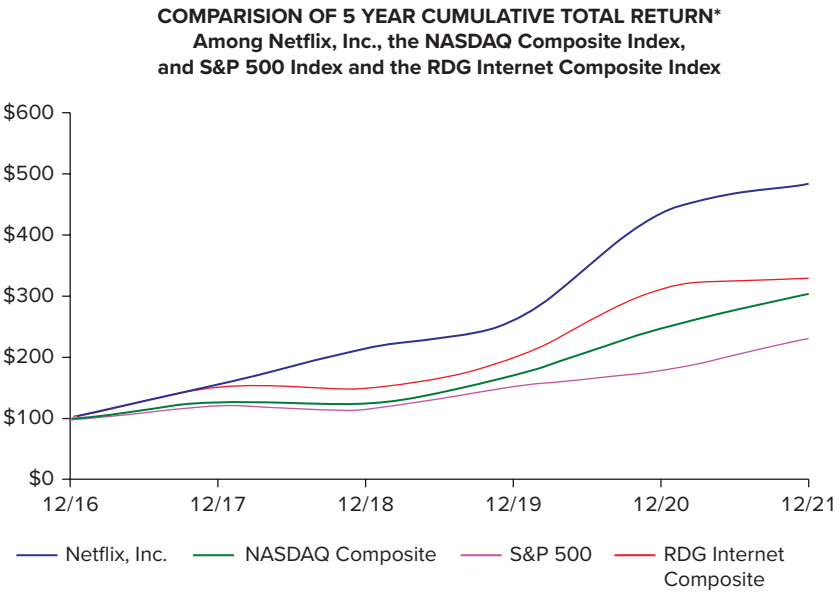
Line Graphs to Compare Performance

Exhibit 9.29 provides an example of reporting as shown in **Netflix**'s Annual Report/10-K, which uses a visualization to summarize the returns to an investment in Netflix against a peer group, the NASDAQ composite, and the S&P 500. The visualization is a line graph that communicates the firm's performance relative to three benchmarks and is particularly useful to current and potential shareholders. We note that Netflix returns more than a $100 more on a theoretical investment made as of 12/16 (December 2016) than any of the three benchmarks, suggesting it would have been a superior financial investment over the time period shown in the graph.

Research Reports and Earnings Forecasts

Financial analysts often prepare research reports that provide important summary information about the companies they evaluate. To prepare these reports, analysts synthesize and forecast financial statements, listen to conference calls, and talk to managers of the company and its competitors. Generally, analysts also give an opinion on whether to buy, hold, or sell a stock. These reports often present tables and graphs of projected stock price and target price range, offering a professional, sophisticated, and useful view of company prospects. Exhibit 9.30 provides an example of *Value Line*'s financial analyst report for **Tesla**.

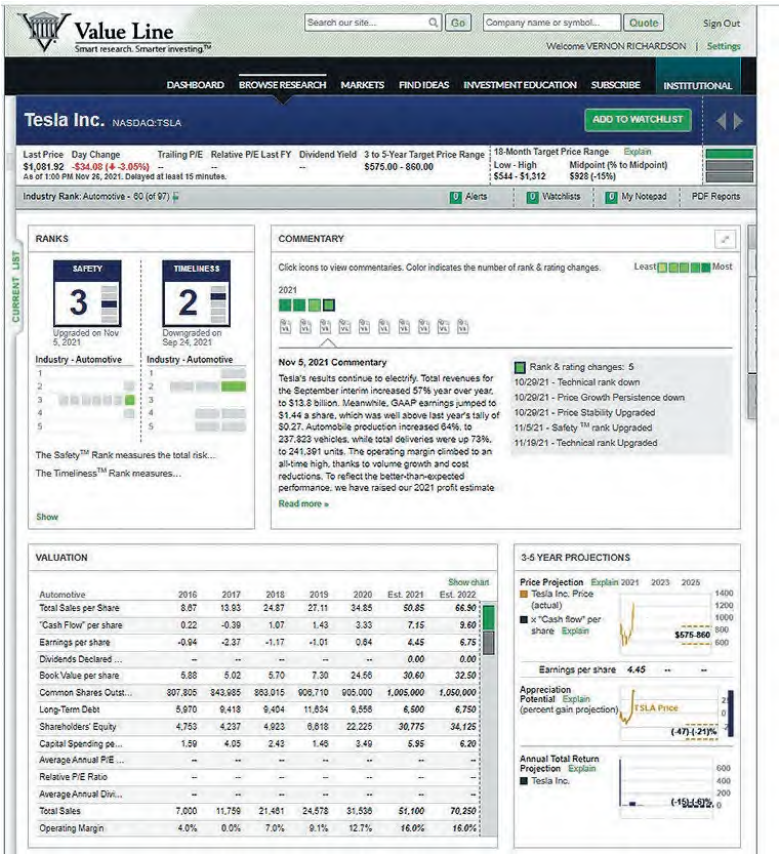
Exhibit 9.29
Netflix Stock Price Performance, 2016–2021



*$100 invested on 12/31/16 in stock or index, including reinvestment of dividends. Fiscal year ending December 31.

Standard & Poor's, a division of S&P Global.

Exhibit 9.30
***Value Line* Financial Analyst Report for Tesla**
(Source: *Value Line*.)



Line Graphs in Sensitivity Analysis

Recall that sensitivity analysis evaluates outcomes based on uncertainty regarding an input. Exhibit 9.31 is a line graph showing how changing assumptions about sales growth and the discount rate used to discount subsequent cash flows affect the prediction of the value of a company's stock price. We can see that a relatively small increase in the discount rate has dramatic implications on the market valuation. We can also see that small adjustments in the sales growth rate appears to have minimal implications for market valuation.

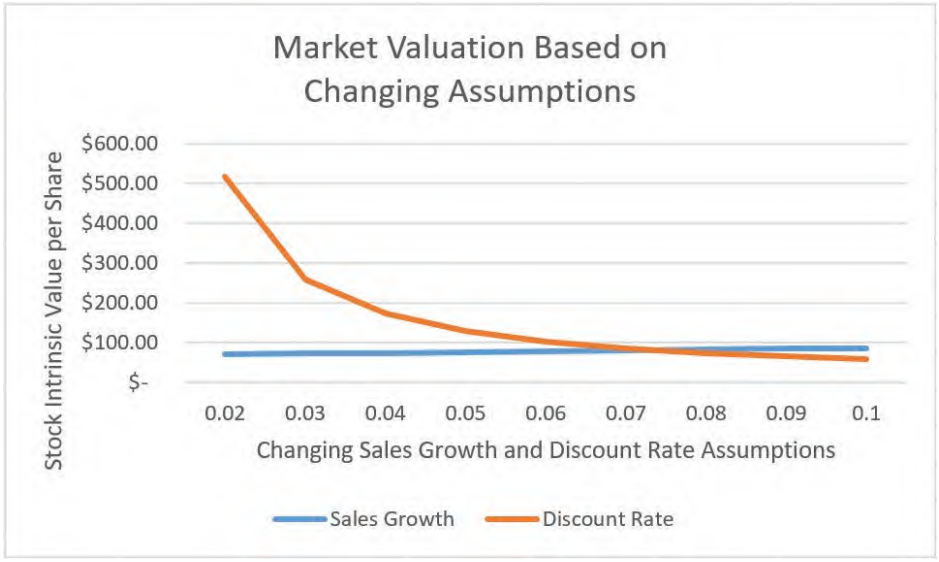


Exhibit 9.31
Sample Sensitivity Analysis Shown in a Line Graph

PROGRESS CHECK

17. Exhibit 9.29 shows Netflix's stock performance relative to three composite market indexes. When did the bulk of the changes occur: early in the time period or in the more recent years?
18. In the sensitivity analysis displayed in Exhibit 9.31, which change of 1 percent (or 0.01) in either sales growth or discount rate is associated with the greater change in the predicted value of a company's stock? How can you tell?

Summary

Data analytics is used in finance to address financial questions. We specified the types of questions addressable with data and how they might differ in the three main branches of finance: corporate finance, investments, and financial markets and institutions. Next, we discussed the data available to address finance questions. We also provided an overview of the different types of analysis that can be performed, including descriptive, diagnostic, predictive, and prescriptive analysis. We concluded with examples of the types of visualizations that can be used to report the results of financial analytics to decision-makers.

Key Terms

anomaly (627) Something that deviates from the expected or the norm.

asset turnover (628) Sales divided by total assets.

capital asset pricing model (CAPM) (633) A model describing the relationship between systematic risk and expected return for financial assets, particularly stocks.

capital budgeting (642) The process of evaluating potential (typically large) investments to help identify which, if any, investments a company should choose.

corporate finance (620) The business function that helps corporations manage their funding sources, capital structure, and investment decisions with the goal of increasing the company's value for its shareholders.

cost of capital (639) The cost of available funds for a company; incorporates the level of risk and acts as a hurdle rate that the company must overcome to create value for its investors.

credit score (636) A statistic that evaluates a customer's creditworthiness and ability to repay a loan.

discount (interest) rate (639) An interest rate used to evaluate an investment based on the risk inherent in the investment.

DuPont ratio analysis (628) Analysis that summarizes and disaggregates company performance into three component ratios to determine the underlying reasons for a company's performance: profit margin, asset turnover, and financial leverage.

EDGAR (624) A repository of all required financial disclosures (including financial statements) submitted to the Securities and Exchange Commission.

equity (628) The value of stockholder investments in a company, calculated as a company's total assets minus total liabilities.

finance (620) The management of money by investing, borrowing, lending, budgeting, saving, and/or forecasting.

financial capital or **capital (620)** Economic resources that companies use to operate their business.

financial institutions (620) Institutions that work in the financial markets, such as banks, investment companies, and brokerages. They generally act as intermediaries, or brokers, that help companies and individuals needing capital (money) find those willing to lend the capital.

financial leverage (628) The extent to which assets are purchased using debt; calculated as total assets divided by stockholders' equity.

financial markets (620) The markets for selling various types of investments, such as stocks and bonds, commodities, and derivatives.

internal rate of return (IRR) (639) The discount rate that makes the project's net present value equal to zero, with the present value of cash outflows equal to the present value of the cash inflows.

investments (branch of finance) (620) The acquisition and disposition of assets with the objective of generating income or profit.

liquidity (620) The availability of cash to meet company obligations.

net present value (NPV) (639) The present value of cash inflows minus the present value of cash outflows.

outlier (627) An observation, or data point, that lies outside its expected distribution.

point of indifference (641) The level at which it does not matter which alternative is chosen because the outcome is the same under either alternative.

profit margin (628) Income divided by total revenues; the extent to which the company earns money on each sales dollar.

ratio analysis (627) Quantitative analysis of financial statement items to help evaluate a company's liquidity, solvency, profitability, and efficiency.

return on assets (ROA) (629) A measure of profitability for each dollar of resources (assets) available for management to use, computed as net income ÷ total assets.

return on equity (ROE) (629) A measure of profitability generated for each dollar of stockholders' investment, defined as net income ÷ stockholders' equity, or net profit (income) ÷ stockholders' equity.

Sharpe ratio (632) The average return to an investment earned in excess of the risk-free rate (on 10-year Treasury bonds) per unit of volatility or total risk.

solvency (620) Financial ability to pay company debts as they come due.

working capital (620) Available capital (such as cash and other resources) that a company can readily use in its day-to-day operations.

XBRL (eXtensible Business Reporting Language) (624) An open, global standard for exchanging financial reporting information.

ANSWERS TO PROGRESS CHECKS

1. Financial markets and institutions is the branch of finance that includes the banks that evaluate potential borrowers and extend loans to them. A bank is an example of a financial institution.
2. Managing a financial portfolio for school teachers' retirement funds takes place in the investments branch of finance. Hedge funds, mutual funds, pension funds, and other companies are all looking for opportunities to invest in financial assets such as stocks of other firms, U.S. Treasury bonds, corporate bonds, and commodities (oil, wheat, gold, and so on).
3. Helping 7-Eleven decide whether to invest in stores in China, Mexico, or Canada would fall to the corporate finance branch of finance, which allocates capital to its most effective opportunity. Generally, those in corporate finance work internally to help company management optimize value for shareholders.
4. If a company is at a significant risk of bankruptcy, you will need to incorporate that risk into your analysis of the potential investment.
5. "What are the expected returns to different financial instruments?" and "What is the chance that a company we might invest in will go bankrupt?" are both examples of finance questions in the investments area that are addressable with financial analytics.
6. "What is the expected level of cash flows in the next month?" and "Will a potential borrower repay a loan?" are both examples of predictive financial analytics.
7. XBRL is the computer-based standard used to efficiently exchange financial information between disclosing companies (publicly traded companies that are required to submit information to the Securities and Exchange Commission) and those who wish to access that information. XBRL allows interested parties to quickly retrieve accurate financial data.
8. While XBRL tags each number that appears in a company's SEC filing in a computerized format to make it easy to access, the SEC filings held in the EDGAR repository are the actual text from the filings (see Exhibit 9.3).
9. A bar graph that shows a company's stock-market returns over the past five years will display the company's financial performance and help to answer the question "What happened?"
10. An analysis of mean, median, standard deviation, and other descriptive statistics regarding total loans outstanding, loans at risk (those that are behind on interest or principal payments), and the distribution of dates that the loans mature would all help to describe a bank's loan portfolio.
11. DuPont analysis decomposes performance (return on equity) into three component ratios: profit margin (profitability), asset turnover, and financial leverage. One diagnostic analytics technique is to drill down into the data to find unknown relationships or

patterns, and DuPont analysis does exactly that: It is a drill-down analysis of return on equity (or return on assets).

12. The Sharpe ratio is a ratio of the average return earned in excess of the risk-free rate (on 10-year Treasury Bonds) to a unit of volatility or total risk. In other words, it is a measure of how much investors are compensated for the extra risk they assume.
13. Predicting bankruptcy risk is important for potential investors and lenders, who do not want to invest in or lend money to a company that is likely to go bankrupt. Predictive analysis is used to help assess a company's risk of bankruptcy.
14. All other things equal, the larger the amount of a requested loan, the lower the chance of repaying the loan. If the prospective borrower requests a large loan, there is a greater likelihood that the loan will not be granted.
15. Prescriptive analytics addresses the question "What should we do, based on what we expect will happen?" Predictive analytics predicts the amount and timing of future cash flows. Prescriptive analytics takes predictive analytics a step further by evaluating different scenarios and levels of risk (and their related discount rates) to help a company decide which action(s) to take.
16. All other things equal, you should make an investment if the internal rate of return (IRR) is greater than the cost of the funds. Here, the IRR (5 percent) is less than the cost of the funds (6 percent), so you should forego the investment.
17. The bulk of the changes occurred later in the five-year period, with greater separation between Netflix and other benchmarks occurring in the more recent years.
18. According to Exhibit 9.31, the discount rate seems to have a greater impact on firm valuation than sales growth does.



Multiple-Choice Questions

1. (LO 9.7) Determining the sensitivity of a bank's profits to changing interest rates is an example of which type of financial analytics?
   a. Prescriptive analytics
   b. Descriptive analytics
   c. Diagnostic analytics
   d. Predictive analytics
2. (LO 9.1) The three primary branches of finance include all of the following except
   a. corporate finance.
   b. investments.
   c. financial market and institutions.
   d. banking.
3. (LO 9.5) Determining how well a company performed relative to its budget is an example of which type of financial analytics?
   a. Prescriptive analytics
   b. Descriptive analytics
   c. Diagnostic analytics
   d. Predictive analytics

4. (LO 9.2) Which branch of finance determines how well the company performed relative to the budget?
   a. Corporate finance
   b. Investments
   c. Financial market and institutions
   d. Banking
5. (LO 9.3) XBRL is a means of disclosing financial information in a computer-readable format. Its submission is required by the
   a. New York Stock Exchange.
   b. Securities and Exchange Commission.
   c. Financial Executives Institute.
   d. Financial Accounting Standards Board.
6. (LO 9.5) The DuPont ratio analysis decomposes return on equity into how many component parts?
   a. Four
   b. Three
   c. Two
   d. Five
7. (LO 9.3) EDGAR serves as a repository for information about
   a. stock prices.
   b. interest rates.
   c. the economy.
   d. financial statements.
8. (LO 9.5) The ratio of the average return earned in excess of the risk-free rate to a unit of volatility or total risk is the
   a. Sharpe ratio.
   b. return on equity (ROE).
   c. return on asset (ROA).
   d. current ratio.
9. (LO 9.4) Which of the following has a higher risk and higher return than a corporate bond?
   a. Treasury bond
   b. Treasury bill
   c. Share of stock
   d. Deposit at a bank
10. (LO 9.6) To predict whether firms are at high or low risk of bankruptcy, analysts can use the predictive analytics technique of
    a. sensitivity analysis.
    b. regression analysis.
    c. classification analysis.
    d. time series analysis.
11. (LO 9.8) Visualizations that show performance relative to relevant benchmarks are examples of
    a. prescriptive analytics.
    b. descriptive analytics.
    c. diagnostic analytics.
    d. predictive analytics.

Discussion Questions

1. (LO 9.1) Compare and contrast the three primary branches of finance.
2. (LO 9.2) What types of questions are representative of those in the investments finance function?
3. (LO 9.3) Describe two sources of financial statement data.
4. (LO 9.4) Explain what the Sharpe ratio tells the analyst.
5. (LO 9.5) How does an analysis of financial ratios, including the DuPont ratios, qualify as diagnostic analysis?
6. (LO 9.6) Predicting whether a loan applicant will receive a loan is an example of predictive analytics. Why is predicting whether the borrower will repay the loan fundamental to that analysis?
7. (LO 9.6) Altman's fifth factor in predicting bankruptcy (X_5) is sales divided by total assets. The greater the sales divided by assets, the greater the possibility of not going bankrupt. Why?
8. (LO 9.6) If a potential borrower has a high amount of debt relative to income, is it more likely or less likely that a loan will be granted to that borrower? Explain.
9. (LO 9.7) How is net present value (NPV) different from the internal rate of return (IRR) in cash flow analysis? How are they similar?
10. (LO 9.8) See Exhibit 9.31. Why is a visualization of sensitivity analysis easier to interpret than a table of sensitivity analysis?

connect

Brief Exercises

1. (LO 9.1, 9.2) Match the branch of finance (corporate finance, investments, or financial markets and institutions) to the question that it answers.

Example Finance Question	**Branch of Finance** (Corporate Finance, Investments, or Financial Markets and Institutions)
Should we expand domestically or internationally? Which will have the greater return?	
Which has higher risk-adjusted returns, stocks or bonds?	
If interest rates go up five percent, how will banking revenues be affected?	
How well did the company perform relative to budget?	
What is the chance that the company in which we are investing will go bankrupt?	
Should our bank extend a loan to a potential borrower?	
Should we make the product ourselves or outsource manufacturing to another company?	

2. (LO 9.4, 9.5, 9.6, 9.7) Match the analytics type (descriptive, diagnostic, predictive, or prescriptive) to the finance question it answers.

Example Finance Question	**Analytics Type** (Descriptive, Diagnostic, Predictive, or Prescriptive)
How well did the company perform last month?	
What is the relationship between the returns earned and the risk borne for financial investments?	
What happens to **General Motors**' profits if the cost of steel increases by 10 percent to 30 percent?	
What is the expected level of company sales next year?	
In which asset should **Federal Express** invest to earn greater profits and be more sustainable: a new fuel-efficient fleet or a new logistics system (which will reduce the number of miles driven)?	
Why do gold prices increase when the economy is struggling?	
What is the sum of real-estate sales in Connecticut?	

3. (LO 9.4, 9.5, 9.6, 9.7) Match the analytics technique to the analytics type (descriptive, diagnostic, predictive, or prescriptive).

Analytics Technique	**Analytics Type** (Descriptive, Diagnostic, Predictive, or Prescriptive)
Cash flow analysis	
Bankruptcy classification	
Drill-down analysis of ROE using DuPont ratios	
Drill-down analysis of ROE using the Sharpe ratio	
Descriptive statistics of stock-market returns	
Sensitivity analysis of **Goldman Sachs**, profits to changes in interest rates	
Time series analysis of future earnings	

4. (LO 9.4, 9.5, 9.6, 9.7) Match the analytics type (descriptive, diagnostic, predictive, or prescriptive) and branch of finance (corporate finance, investments, or financial markets and institutions) to the analytics technique used.

Analytics Technique	**Analytics Type** (Descriptive, Diagnostic, Predictive, or Prescriptive)	**Branch of Finance** (Corporate Finance, Investments, or Financial Markets and Institutions)
Classification to decide who will receive a loan and who will not		
Time series analysis to forecast sales		
Drill-down analysis of stock returns and corporate bonds using the Sharpe ratio		
Regression analysis describing determinants of return in the stock market or corporate bond market		
Drill-down analysis of the analyst's own company		
Summary of the performance of the banking industry in the United States		
Calculation of net present value (NPV) and internal rate of return (IRR) for a company's potential capital investments		

5. (LO 9.4, 9.5, 9.6, 9.7) Match the analytics type (descriptive, diagnostic, predictive, or prescriptive) to the financial analytics techniques used.

Financial Analytics Technique	**Analytics Type** (Descriptive, Diagnostic, Predictive, or Prescriptive)
Ratio analytics	
Performance different than expected	
Scenario analytics	
Pivot table analytics to find the most profitable products or customers	
Finding unexpected patterns in performance	
Cash-flow analysis	
Goal-seek analysis	

6. (LO 9.4) Match each finance question to the descriptive analytics statistical technique (count, histogram, ratio analysis, sum, or standard deviation).

Finance Question	**Data Analytics Statistical Technique** (Count, Histogram, Ratio Analysis, Sum, or Standard Deviation)
How many banks are in Texas?	
How can we evaluate risk?	
What are the total profits for the retail sector?	
What is the frequency of bankruptcy in the retail sector by year?	
What is the price-to-earnings ratio for each company in the industry?	

Problems

1. (LO 9.5) Compute the DuPont ratios (profit margin, asset turnover, and financial leverage ratios) for **Walmart** (WMT) and **Target** (TGT), given the following numbers for 2012–2020. Check your numbers to make sure that Profit margin × Asset turnover × Financial leverage = Return on equity. The data file, DuPont Analysis Walmart Target Data.xlsx, is available in Connect or via the Additional Student Resources page. (*Note:* The data are in $ millions.)[4]

WMT				
Year	**Sales Revenue**	**Net Income**	**Assets**	**Stockholders' Equity**
2020	$ 559,151	$13,510	$ 252,496	$ 87,531
2019	$ 523,964	$14,881	$ 236,495	$ 81,552
2018	$ 514,343	$ 6,670	$ 219,295	$ 79,634
2017	$ 500,343	$10,523	$ 204,522	$ 80,822
2016	$ 485,873	$14,293	$ 198,825	$ 80,535
2015	$ 482,130	$15,080	$ 199,581	$ 83,611
2014	$ 485,651	$17,099	$ 203,490	$ 85,937
2013	$ 476,294	$16,695	$ 204,541	$ 81,339
2012	$ 468,651	$17,756	$ 203,105	$ 81,738

[4]Source: XBRL (Calcbench) from company financial statements.

TGT				
Year	Sales Revenue	Net Income	Assets	Stockholders' Equity
2020	$93,561	$4,368	$54,248	$14,440
2019	$78,112	$3,281	$42,779	$11,833
2018	$75,356	$2,937	$41,290	$11,297
2017	$71,879	$2,914	$40,303	$11,651
2016	$69,495	$2,737	$37,431	$10,915
2015	$73,785	$3,363	$40,262	$12,957
2014	$72,618	$(1,636)	$41,172	$13,997
2013	$71,279	$1,971	$44,553	$16,231
2012	$73,301	$2,999	$48,163	$16,558

2. (LO 9.5) Compute the DuPont ratios (profit margin, asset turnover, and financial leverage ratios) for **Amazon** (AMZN) and **Kroger** (KR), given the following numbers for 2012–2020. Check your numbers to make sure that Profit margin × Asset turnover × Financial leverage = Return on equity. The data file, DuPont Analysis Amazon Kroger Data.xlsx, is available in Connect or via the Additional Student Resources page. (*Note:* The data are in $ millions.)[5]

AMZN				
Year	Sales Revenue	Net Income	Assets	Stockholders' Equity
2020	$386,064	$21,331	$321,195	$93,404
2019	$280,522	$11,588	$225,248	$62,060
2018	$232,887	$10,073	$162,648	$43,549
2017	$177,866	$ 3,033	$131,310	$27,709
2016	$135,987	$ 2,371	$ 83,402	$19,285
2015	$107,006	$ 596	$ 64,747	$13,384
2014	$ 88,988	$ (241)	$ 54,505	$10,741
2013	$ 74,452	$ 274	$ 40,159	$ 9,746
2012	$ 61,093	$ (39)	$ 32,555	$ 8,192

KR				
Year	Sales Revenue	Net Income	Assets	Stockholders' Equity
2020	$132,498	$2,556	$48,256	$9,576
2019	$122,286	$1,640	$45,256	$8,602
2018	$121,162	$3,076	$38,118	$7,886
2017	$122,662	$1,889	$37,197	$6,905
2016	$115,337	$1,957	$36,505	$6,710
2015	$109,830	$2,049	$33,897	$6,798
2014	$108,465	$1,747	$30,497	$5,442
2013	$ 98,375	$1,531	$29,281	$5,395
2012	$ 96,619	$1,508	$24,634	$4,214

[5]Source: XBRL (Calcbench) from company financial statements.

3. (LO 9.6) Analysts use Altman's Z to predict bankruptcy.

Required:

1. According to Altman's Z, which of the following factors help predict bankruptcy?

Factor	Predicts Bankruptcy? (Yes/No)
Working capital ÷ Total assets	
Return on equity	
Sales ÷ Total assets	
Stock market performance over the past year	
Sales growth	
Competitor profit margin	
Depreciation expense	
Retained earnings ÷ Total assets	

4. Classify each Altman's Z-score as either in the distress zone, gray zone, or safe zone with respect to risk of bankruptcy.

Altman's Z-Score	Distress Zone, Gray Zone, or Safe Zone
1.81	
3.05	
2.57	
1.03	
5.73	
–0.25	

5. (LO 9.6) List five factors (independent variables) that a bank might use to determine whether a prospective borrower will be able to repay a loan. Is each of these factors positively or negatively related to a loan approval? Explain.
6. (LO 9.4, 9.5) Which has the highest return on investment: the S&P 500, Baa (high-grade) corporate bonds, or U.S. Treasury bonds? Download the Investment Returns data set from Connect or Additional Student Resources. (Source: http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/histretSP.html.)

   Required

   1. Calculate the average returns for the S&P 500 stock index, Baa (high-grade) corporate bonds, and U.S. Treasury bonds from 1928 to 2021. Which has the highest return?
   2. Calculate the standard deviation of returns to the S&P 500 stock index, Baa (high-grade) corporate bonds, and U.S. Treasury bonds from 1928 to 2021. Which has the highest standard deviation of returns? (*Hint:* Use the STDEV.S Excel function for each index.)
   3. Generally we expect that the higher the risk, the greater the expected return. Which investment has the highest risk (measured here as standard deviation of returns), and which has the greatest return?

7. (LO 9.4, 9.5) Which has the highest return on investment: the S&P 500, high-grade corporate bonds, or U.S. Treasury bonds? Download the Investment Returns data

set from Connect or Additional Student Resources. (Source: http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/histretSP.html.)

Required

1. Using the Data Analysis ToolPak in Excel, perform a *t*-test: Two-Sample Assuming Equal Variances between the annual returns on the S&P 500 stock index and Baa (high-grade) corporate bonds. (*Hint:* Set Hypothesized Mean Difference to 0 (zero) and alpha at 0.05.) What *p*-value is associated with the *t*-test (using a one-tailed test)? Is that *p*-value less than our established alpha of 0.05? Is one *p*-value statistically greater than the other (using a one-tailed test)?
2. Using the Data Analysis ToolPak in Excel, perform a *t*-test: Two-Sample Assuming Equal Variances between the annual returns on the high-grade corporate bonds and U.S. Treasury bonds. (*Hint:* Set Hypothesized Mean Difference to 0 (zero) and alpha at 0.05.) What *p*-value is associated with the *t*-test? Is that *p*-value less than our established alpha of 0.05? Is one *p*-value statistically greater than the other (using a one-tailed test)?
3. What can you conclude about the returns to these different investments?
4. Given the results above, what would you expect to be the result of a *t*-test between the S&P 500 stock index and U.S. Treasury bonds?

8. Different industries have different levels of returns. This problem works to compare industries to each other, including gross margin, net profit margin, and R&D expenditures as a percentage of total sales. Download the Margin data set from Connect or Additional Student Resources. (Source: https://pages.stern.nyu.edu/~adamodar/.)

Required

1. Calculate the highest net margin (net profit margin) for all industries.
2. Which industry had the highest gross margin? The lowest gross margin? How can you explain the industry with the highest gross margin? (*Hint:* What would be considered in cost of goods sold?)
3. Which industry spent the most on R&D expenditures as a percentage of sales? Is it what you expected? Explain.

9. The lower the cost of capital, the better for a company. The cost of capital is the cost to use someone else's money or investment such as the interest rate of the debt or the cost of shareholder's investment (or equity). Which industries have the lowest cost of capital? Download the WACC data set from Connect or Additional Student Resources. (Source: https://pages.stern.nyu.edu/~adamodar/.)

Required

1. Column D has the cost of equity, which you can think of as the cost that shareholders expect to invest in equity in companies in that industry. What is average cost of equity for all industries? Which industry has the highest cost of equity? Why do you think that might be?
2. Column G has the interest rate (or cost) of debt, which you can think of as the interest rate on debt. What is the average cost of debt for all industries? Which industry has the highest rate of debt? Why do you think that might be?
3. Column K has the combined or weighted cost of capital. What is the average weighted cost of capital for all industries? Which industry has the highest cost of capital? Why do you think that might be?

Data Visualization Exercises



DATA VISUALIZATION

1. (LO 9.4) Download the sp-500-historical-annual-returns data set in Excel from Connect or Additional Student Resources.[6] Note the annual stock market returns to investment in the S&P 500. Create a visualization in Excel using a bar chart. Note the volatility of returns over the years, which is an indicator that the stock market is a risky investment.

[6]Source: https://www.macrotrends.net/2526/sp-500-historical-annual-returns.

Required

1. Create a bar chart of the annual returns (*Hint:* **Insert > Charts > Bar Chart**) or column chart of the annual returns (*Hint:* **Insert > Charts > Column Chart**).
2. Label the vertical axis as "Annual Percentage Return", the horizontal axis as "Year", and the chart title as "Annual Returns to the S&P 500 Index".

2. (LO 9.4) Download the nasdaq-by-year-historical-annual-returns data set in Excel from Connect or Additional Student Resources.[7] Note the annual stock market returns to investment in the NASDAQ stock market index. Use any type of visualization (such as histograms, bar charts, or column charts) that will allow you to complete the following requirements.

Required

1. Create a chart of the annual returns. (*Hint:* **Insert > Charts**.) Submit this screenshot to your instructor if assigned.
2. What is the average return on the NASDAQ market? The median return? (*Hint:* Use AVERAGE() and MEDIAN() functions in Excel.)
3. Based on the visualization, how many annual returns were 20 percent or higher? (*Hint:* Perform a count.)
4. Based on the visualization, how many annual returns were negative? (*Hint:* Perform a count.)

[7]Source: https://www.macrotrends.net/2623/nasdaq-by-year-historical-annual-returns.

LAB 9.1 **Excel:** Descriptive and Diagnostic Analytics: Calculating Returns to Investments Using the Sharpe Ratio
LAB 9.2 **Excel:** Diagnostic Analytics: Applying DuPont Analysis of Financial Performance
LAB 9.3 **Excel:** Predictive Analytics: Evaluating Loan Acceptance
LAB 9.4 **Excel:** Predictive Analytics: Predicting Bankruptcy
LAB 9.5 **Excel:** Prescriptive Analytics: Evaluating Investments Using NPV
LAB 9.6 **Excel:** Prescriptive Analytics: Evaluating Investments Using IRR

Lab 9.1 Excel

Descriptive and Diagnostic Analytics: Calculating Returns to Investments Using the Sharpe Ratio

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Descriptive Statistics, Investment Returns

Decision-Making Context

Descriptive analytics uses facts and statistics to explain what happened. In this lab, we report descriptive statistics on the returns to three different investments: the S&P 500 stock index, high-quality corporate bonds (Baa rating), and a U.S. Treasury bond.

We will also use the Sharpe ratio to incorporate the return and relative risk to these financial investments.

Required

1. Compute descriptive statistics for the three financial investments from 1928 to 2021.
2. Compute the Sharpe ratio for excess returns.

Specify the Question

What are the descriptive statistics for the returns to various financial investments? What are the risks relative to the returns for various financial instruments as measured by the Sharpe ratio?

Obtain the Data

The reported results of financial performance are included in the data set for three different investments.

Data Dictionary

Year:	Year between 1928 and 2021
S&P 500 (includes dividends):	The annual return to the S&P 500 (including dividends)
Baa Corporate Bond:	The annual return to a Baa-rated (high-quality) corporate bond. Baa represents a relatively low-risk bond.
US T. Bond:	The annual return to the 10-year U.S. Treasury bond.

Analyze the Data

Open Excel file Lab 9.1 Data.xlsx. Lab Exhibit 9.1.1 provides a snippet of the spreadsheet.

Lab Exhibit 9.1.1

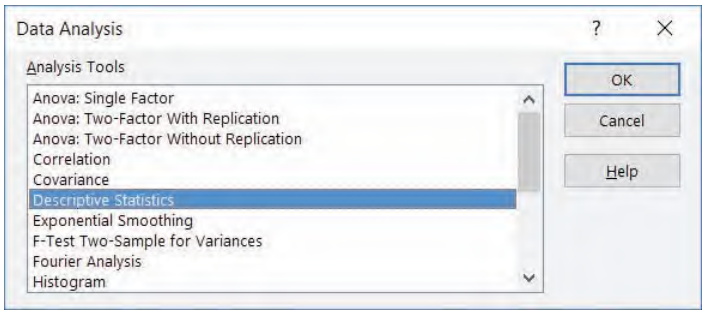
	A	B	C	D
1	*Year*	*S&P 500 (includes dividends)*	*Baa Corporate Bond*	*US T. Bond*
2	1928	43.81%	3.22%	0.84%
3	1929	-8.30%	3.02%	4.20%
4	1930	-25.12%	0.54%	4.54%
5	1931	-43.84%	-15.68%	-2.56%
6	1932	-8.64%	23.59%	8.79%
7	1933	49.98%	12.97%	1.86%
8	1934	-1.19%	18.82%	7.96%
9	1935	46.74%	13.31%	4.47%
10	1936	31.94%	11.38%	5.02%
11	1937	-35.34%	-4.42%	1.38%
12	1938	29.28%	9.24%	4.21%
13	1939	-1.10%	7.98%	4.41%
14	1940	-10.67%	8.65%	5.40%
15	1941	-12.77%	5.01%	-2.02%
16	1942	19.17%	5.18%	2.29%
17	1943	25.06%	8.04%	2.49%
18	1944	19.03%	6.57%	2.58%
19	1945	35.82%	6.80%	3.80%
20	1946	-8.43%	2.51%	3.13%
21	1947	5.20%	0.26%	0.92%

We will now learn how to get the descriptive statistics for each of the three investments from 1928 to 2021.

Analysis Task 1: Calculate Summary Statistics for Each of the Three Investments.

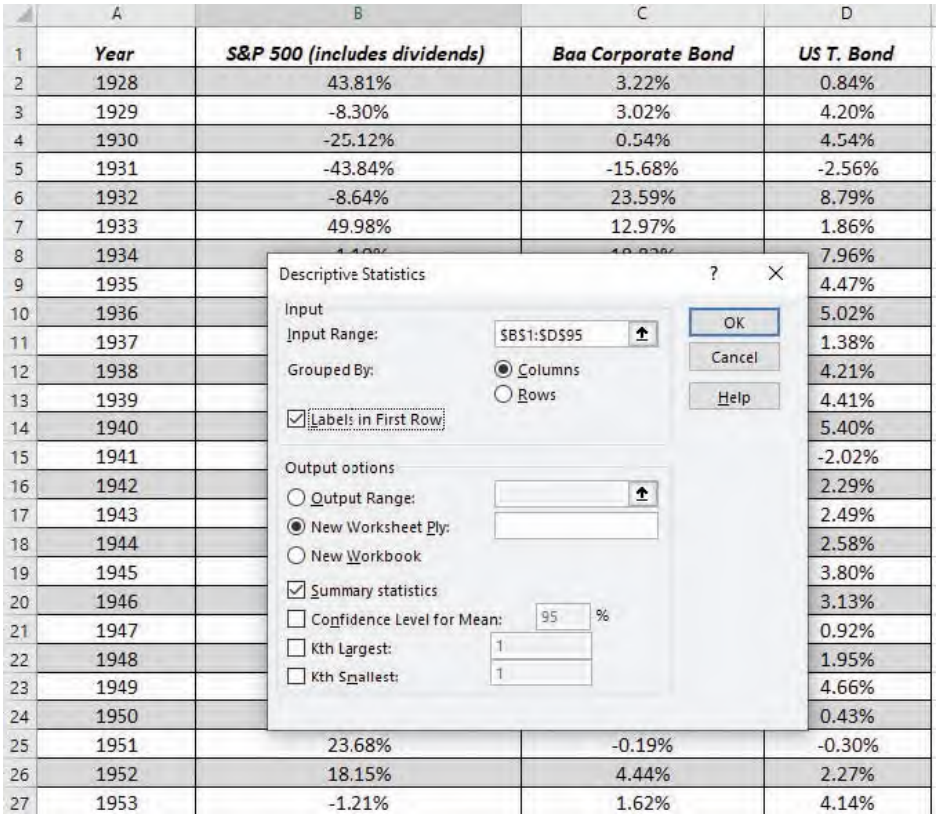
1. To run the descriptive statistics, we need to make sure the Data Analysis ToolPak is loaded. Look at the **Data > Analysis** and see if the Data Analysis add-in has been installed. If you haven't installed the Data Analysis ToolPak, textbook Appendix E provides instructions for installing it.
2. To perform the descriptive statistics, select **Data > Analysis > Data Analysis.** A dialog box will open. Select Descriptive Statistics and click OK (Lab Exhibit 9.1.2).

Lab Exhibit 9.1.2



3. We will first compute the summary statistics for each variable. For the Input Range, select cells $B$1:$D$95 (including the titles).
4. Check Labels in First Row to show that we have labels in the first row.
5. Check the box next to Summary statistics and click OK. The completed dialog box appears as shown in Lab Exhibit 9.1.3.

Lab Exhibit 9.1.3



	A	B	C	D
1	*Year*	*S&P 500 (includes dividends)*	*Baa Corporate Bond*	*US T. Bond*
2	1928	43.81%	3.22%	0.84%
3	1929	-8.30%	3.02%	4.20%
4	1930	-25.12%	0.54%	4.54%
5	1931	-43.84%	-15.68%	-2.56%
6	1932	-8.64%	23.59%	8.79%
7	1933	49.98%	12.97%	1.86%
8	1934			7.96%
9	1935			4.47%
10	1936			5.02%
11	1937			1.38%
12	1938			4.21%
13	1939			4.41%
14	1940			5.40%
15	1941			-2.02%
16	1942			2.29%
17	1943			2.49%
18	1944			2.58%
19	1945			3.80%
20	1946			3.13%
21	1947			0.92%
22	1948			1.95%
23	1949			4.66%
24	1950			0.43%
25	1951	23.68%	-0.19%	-0.30%
26	1952	18.15%	4.44%	2.27%
27	1953	-1.21%	1.62%	4.14%

The resulting worksheet tab is as shown in Lab Exhibit 9.1.4. We note the minimum and maximum for each of the three investments.

Lab Exhibit 9.1.4

	A	B	C	D	E	F
1	*S&P 500 (includes dividends)*		*Baa Corporate Bond*		*US T. Bond*	
2						
3	Mean	0.118206	Mean	0.071857	Mean	0.051099
4	Standard Error	0.020072	Standard Error	0.007748	Standard Error	0.007921
5	Median	0.1452	Median	0.06465	Median	0.0328
6	Mode	#N/A	Mode	#N/A	Mode	#N/A
7	Standard Deviation	0.194608	Standard Deviation	0.075121	Standard Deviation	0.076795
8	Sample Variance	0.037872	Sample Variance	0.005643	Sample Variance	0.005897
9	Kurtosis	0.125039	Kurtosis	0.64706	Kurtosis	1.642672
10	Skewness	-0.44474	Skewness	0.296424	Skewness	0.980655
11	Range	0.964	Range	0.4473	Range	0.4393
12	Minimum	-0.4384	Minimum	-0.1568	Minimum	-0.1112
13	Maximum	0.5256	Maximum	0.2905	Maximum	0.3281
14	Sum	11.1114	Sum	6.7546	Sum	4.8033
15	Count	94	Count	94	Count	94

You'll note that the average return to the S&P is higher than the return to the Baa corporate bond, consistent with the risk/return trade-off, which suggests that stock returns (S&P 500) should be higher than the returns to the corporate bond. And both have greater returns than the risk-free Treasury bond.

So, the average returns follow this order:

S&P 500 returns (0.118) > Baa corporate bond returns (0.072) > US T. bond returns (0.051)

What is the risk of these investments? One way to measure the risk is the standard deviation of these annual returns. Generally, the higher the standard deviation, the greater the risk. Note that the standard deviation of the S&P 500 stock returns is greater than the standard deviation of the Baa corporate bond and the U.S. Treasury bond:

S&P 500 risk (standard deviation) > Baa corporate bond and U.S. Treasury bond risk (standard deviation)



DATA VISUALIZATION

6. To more easily see the risk, create a bar chart of annual S&P 500 returns in one visualization (**Insert > Charts > Statistic Chart > Bar Chart**) by highlighting the S&P 500 returns data. Create a bar chart of annual Baa returns in another visualization by highlighting the Baa corporate returns data. Then compare and contrast both the mean and the standard deviation of returns.

Analysis Task 2: Calculate the Sharpe Ratio.

The next requirement is to compute the Sharpe ratio. The Sharpe ratio is a widely used method for calculating the risk-adjusted return. We already know that the return is greater for the S&P 500 than the return for the corporate bond. We also know that the standard deviation is greater for the S&P 500, suggesting that its risk is greater.

The Sharpe ratio is computed as the average return earned in excess of the risk-free rate (on the U.S. Treasury bond).

1. In cell F1, insert the title "S&P Excess".
2. In cell G1, insert the title "Baa Corporate Excess".
3. In cell F2, insert the formula =B2-D2. Copy the result from cell F2 to cells F3:F95.
4. In cell G2, insert the formula =C2-D2. Copy the result from cell G2 to cells G3:G95. The result is as shown in Lab Exhibit 9.1.5.

Lab Exhibit 9.1.5

	A	B	C	D	E	F	G
1	*Year*	*S&P 500 (includes dividends)*	*Baa Corporate Bond*	*US T. Bond*		*S&P Excess*	*Baa Corporate Excess*
2	1928	43.81%	3.22%	0.84%		42.97%	2.38%
3	1929	-8.30%	3.02%	4.20%		-12.50%	-1.18%
4	1930	-25.12%	0.54%	4.54%		-29.66%	-4.00%
5	1931	-43.84%	-15.68%	-2.56%		-41.28%	-13.12%
6	1932	-8.64%	23.59%	8.79%		-17.43%	14.80%
7	1933	49.98%	12.97%	1.86%		48.12%	11.11%
8	1934	-1.19%	18.82%	7.96%		-9.15%	10.86%
9	1935	46.74%	13.31%	4.47%		42.27%	8.84%
10	1936	31.94%	11.38%	5.02%		26.92%	6.36%
11	1937	-35.34%	-4.42%	1.38%		-36.72%	-5.80%
12	1938	29.28%	9.24%	4.21%		25.07%	5.03%
13	1939	-1.10%	7.98%	4.41%		-5.51%	3.57%
14	1940	-10.67%	8.65%	5.40%		-16.07%	3.25%
15	1941	-12.77%	5.01%	-2.02%		-10.75%	7.03%

5. In cell D99, insert the title "Avg. (Excess) Return". This is the return beyond the risk-free rate.
6. In cell D100, insert the title "Standard Deviation".
7. In cell D101, insert the title "Sharpe Ratio".
8. In cell F98, insert the title "S&P 500".
9. In cell G98, insert the title "Baa Corporate Bonds".
10. In cell F99, insert the formula "=AVERAGE(F2:F95)" to calculate the average return of the S&P 500.
11. In cell F100, insert the formula "=STDEV(F2:F95)" to calculate the standard deviation of the S&P 500.
12. In cell F101, insert the formula "=F99/F100" to calculate the Sharpe ratio for the S&P 500.
13. In cell G99, insert the formula "=AVERAGE(G2:G95)" to calculate the average return of the Baa corporate bonds.

14. In cell G100, insert the formula "=STDEV(G2:G95)" to calculate the standard deviation of the Baa corporate bond.
15. In cell G101, insert the formula "=G99/G100" to calculate the Sharpe ratio for the Baa corporate bonds.
16. Format all of the cells as percentages with two digits.

The resulting spreadsheet should appear as shown in Lab Exhibit 9.1.6.

Lab Exhibit 9.1.6

	D	E	F	G
97				
98			S&P 500	Baa Corporate Bonds
99	Avg. (Excess) Return		6.71%	2.08%
100	Standard Deviation		21.09%	6.83%
101	Sharpe Ratio		31.82%	30.40%

The Sharpe ratio for the S&P 500 and Baa corporate bonds reveals a very close relationship between risk and return for both of them, suggesting that the risk is compensated in a similar way by the returns for both the S&P 500 and the Baa corporate bonds.

Report the Results

Often it is smart to share not only the return of an investment but also the amount of risk borne. With the descriptive statistics and the computation of the Sharpe ratio, business analysts are in a position to talk with decision-makers about both return and risk, helping them evaluate potential investments.

Assessment

1. Take a screenshot of your summary statistics, paste it into a Word document named "Lab 9.1 Excel Submission.docx", and label the screenshot Submission 1.
2. Take a screenshot of Sharpe ratio computation, paste it into the same Word document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 9.1 Excel Submission.docx via Connect, if assigned.

There is no alternate lab for Lab 9.1 Excel.

Lab 9.2 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Diagnostic Analytics: Applying DuPont Analysis of Financial Performance

Keywords

Financial Performance, Financial Statement Analysis, Ratio Analysis

Decision-Making Context

The DuPont ratio summarizes and disaggregates company performance and can help answer the important question "What happened?" The DuPont ratio was originally developed by the DuPont Corporation to decompose performance into three component parts, as follows:

$$\begin{aligned}\text{Return on equity (ROE)} &= \text{Profit margin} * \text{Asset turnover} * \text{Financial leverage} \\ &= (\text{Net profit/Sales}) * (\text{Sales/Total assets}^{8}) * \\ &\quad (\text{Total assets/Stockholders' equity}^{8})\end{aligned}$$

DuPont analysis decomposes return on equity into three different ratios: Profit margin (or profitability) ratio, asset turnover ratio, and financial leverage ratio. Business analysts are interested in calculating return on equity because ROE is a measure of profitability generated for each dollar of stockholders' investment. Disaggregating ROE helps analysts understand why ROE increased or decreased, or why the ROE differs between companies.

Required

Perform DuPont analysis on the return on equity (ROE) for the **Hershey** Company and **Smucker's**.

Specify the Question

How can DuPont analysis be used to disaggregate return on equity into profit margin, asset turnover, and financial leverage component ratios?

Obtain the Data

1. Open Excel file Lab 9.2 Data.xlsx.
2. Browse the spreadsheet, which includes sales revenue, net income, and total assets from 2015 to 2020 for both Hershey (Ticker Symbol: HSY) and Smucker's (Ticker Symbol: SJM). A snippet of the spreadsheet appears in Lab Exhibit 9.2.1.

We're now ready to perform the DuPont analysis.

Analyze the Data

Analysis Task 1: Compute the Three Ratios as Components of the DuPont Analysis.

We need to compute three ratios for each company and for each year.

1. To start, recopy the years to a second location in the spreadsheet. Copy Column A (using Ctrl-C) to Column F.
2. In cells G2 and G11, type the words "Profit Margin".
3. In cells H2 and H11, type the words "Asset Turnover".
4. In cells I2 and I11, type the words "Financial Leverage".
5. In cells J2 and J11, type the words "Return on Equity".

[8] Average total assets are often used instead of total assets at the end of the year. We use total assets at the end of the year for our examples in these labs and in the text. In addition, total stockholders' equity at the end of the year is used instead of average stockholders' equity for our examples here and in the text.

Lab Exhibit 9.2.1

	A	B	C	D	E
1	HSY	($ thousands)			
2	Year	Sales Revenue	Net Income	Assets	Stockholder's Equity
3	2020	$8,149,719	$1,278,708	$9,131,845	$2,234,352
4	2019	$7,986,252	$1,149,692	$8,140,395	$1,739,222
5	2018	$7,791,069	$1,177,562	$7,703,020	$1,398,721
6	2017	$7,515,426	$756,537	$5,553,726	$931,565
7	2016	$7,440,181	$720,044	$5,524,333	$827,687
8	2015	$7,386,626	$512,951	$5,344,371	$1,047,462
9					
10	SJM	($ thousands)			
11	Year	Sales Revenue	Net Income	Assets	Stockholder's Equity
12	2020	$8,002,700	$872,600	$16,284,200	$8,124,800
13	2019	$7,801,000	$775,100	$16,970,400	$8,190,900
14	2018	$7,838,000	$511,800	$16,711,300	$7,970,500
15	2017	$7,357,100	$1,338,600	$15,301,200	$7,891,100
16	2016	$7,392,300	$592,300	$15,639,700	$6,850,200
17	2015	$7,811,200	$688,700	$15,984,100	$7,008,500

6. In cell G3, insert "=C3/B3" to compute profit margin for Hershey in 2020, by taking net income and dividing by sales revenue.
7. Copy the contents of cell G3 to all relevant cells in Column G for the profit margin for each respective year for Hershey and for Smucker's.
8. In cell H3, insert "=B3/D3" to compute asset turnover for Hershey in 2020, by taking sales revenue and dividing by assets.
9. Copy the contents of cell H3 to all relevant cells in Column H for the asset turnover for each respective year for Hershey and for Smucker's.
10. In cell I3, insert "=D3/E3" to compute financial leverage for Hershey in 2020, by taking assets and dividing by stockholders' equity.
11. Copy the contents of cell I3 to all relevant cells in Column I for the financial leverage for each respective year for Hershey and for Smucker's.
12. In cell J3, insert "=G3*H3*I3" to compute return on equity for Hershey in 2020, by multiplying all three of the measurements we just calculated.
13. Copy the contents of cell J3 to all relevant cells in Column J for the return-on-equity calculation for each respective year for Hershey and for Smucker's.

The formulas should appear as shown in Lab Exhibit 9.2.2.

Lab Exhibit 9.2.2

	F	G	H	I	J
1	HSY				
2	Year	Profit Margin	Asset Turnover	Financial Leverage	Return on Equity
3	2020	=C3/B3	=B3/D3	=D3/E3	=G3*H3*I3
4	2019	=C4/B4	=B4/D4	=D4/E4	=G4*H4*I4
5	2018	=C5/B5	=B5/D5	=D5/E5	=G5*H5*I5
6	2017	=C6/B6	=B6/D6	=D6/E6	=G6*H6*I6
7	2016	=C7/B7	=B7/D7	=D7/E7	=G7*H7*I7
8	2015	=C8/B8	=B8/D8	=D8/E8	=G8*H8*I8
9					
10	SJM				
11	Year	Profit Margin	Asset Turnover	Financial Leverage	Return on Equity
12	2020	=C12/B12	=B12/D12	=D12/E12	=G12*H12*I12
13	2019	=C13/B13	=B13/D13	=D13/E13	=G13*H13*I13
14	2018	=C14/B14	=B14/D14	=D14/E14	=G14*H14*I14
15	2017	=C15/B15	=B15/D15	=D15/E15	=G15*H15*I15
16	2016	=C16/B16	=B16/D16	=D16/E16	=G16*H16*I16
17	2015	=C17/B17	=B17/D17	=D17/E17	=G17*H17*I17

The spreadsheet should appear as shown in Lab Exhibit 9.2.3.

Lab Exhibit 9.2.3

	A	B	C	D	E	F	G	H	I	J
1	HSY	($ thousands)				HSY				
2	Year	Sales Revenue	Net Income	Assets	Stockholder's Equity	Year	Profit Margin	Asset Turnover	Financial Leverage	Return on Equity
3	2020	$8,149,719	$1,278,708	$9,131,845	$2,234,352	2020	0.1569	0.8925	4.0870	0.5723
4	2019	$7,986,252	$1,149,692	$8,140,395	$1,739,222	2019	0.1440	0.9811	4.6805	0.6610
5	2018	$7,791,069	$1,177,562	$7,703,020	$1,398,721	2018	0.1511	1.0114	5.5072	0.8419
6	2017	$7,515,426	$756,537	$5,553,726	$931,565	2017	0.1007	1.3532	5.9617	0.8121
7	2016	$7,440,181	$720,044	$5,524,333	$827,687	2016	0.0968	1.3468	6.6744	0.8699
8	2015	$7,386,626	$512,951	$5,344,371	$1,047,462	2015	0.0694	1.3821	5.1022	0.4897
9										
10	SJM	($ thousands)				SJM				
11	Year	Sales Revenue	Net Income	Assets	Stockholder's Equity	Year	Profit Margin	Asset Turnover	Financial Leverage	Return on Equity
12	2020	$8,002,700	$872,600	$16,284,200	$8,124,800	2020	0.1090	0.4914	2.0043	0.1074
13	2019	$7,801,000	$775,100	$16,970,400	$8,190,900	2019	0.0994	0.4597	2.0719	0.0946
14	2018	$7,838,000	$511,800	$16,711,300	$7,970,500	2018	0.0653	0.4690	2.0966	0.0642
15	2017	$7,357,100	$1,338,600	$15,301,200	$7,891,100	2017	0.1819	0.4808	1.9390	0.1696
16	2016	$7,392,300	$592,300	$15,639,700	$6,850,200	2016	0.0801	0.4727	2.2831	0.0865
17	2015	$7,811,200	$688,700	$15,984,100	$7,008,500	2015	0.0882	0.4887	2.2807	0.0983



DATA VISUALIZATION

Report the Results

Copy the return on equity for both companies and put them side by side. What is the most effective way to communicate these differences to decision-makers? What is more effective for comparison purposes, a bar chart or a table?

We have now completed the DuPont analysis and can share it with decision-makers to learn how each company performed with respect to profit margin, asset turnover, and financial leverage. This information may be useful to potential investors considering an equity investment, banks considering giving a loan to one or both of the companies, or for the companies themselves to evaluate their own performance.

Assessment:

1. Take a screenshot of the resulting DuPont analysis, paste it into a Word document named "Lab 9.2 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Lab 9.2 Excel Submission.docx via Connect, if assigned.

Alternate Lab 9.2 Excel

Diagnostic Analytics: Applying DuPont Analysis of Financial Performance

Apply the same steps in Lab 9.2 to the Alt Lab 9.2 Data.xlsx data set to perform DuPont analysis for **Kellogg's** and **Kraft Heinz**.

Required

Perform DuPont analysis for Kellogg's and Kraft Heinz.

Open Excel file Alt Lab 9.2 Data.xlsx.

Assessment

1. Take a screenshot of the resulting DuPont analysis, paste it into a Word document named "Alt Lab 9.2 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 9.2 Excel Submission.docx via Connect, if assigned.

Lab 9.3 Excel

Predictive Analytics: Evaluating Loan Acceptance

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Loan Acceptance, Borrower Risk, Borrower Profile

Decision-Making Context

Companies and individuals borrow money for a variety of reasons, such as refinancing other debt and paying off credit cards. Lenders evaluate potential borrowers by evaluating the probability that the borrowers will repay the loan. Based on characteristics of loans that were both granted and not granted, or accepted and rejected, we can better understand lenders' decision-making process.

This lab analyzes the factors that predict loan acceptance by analyzing applicants who were given a loan and applicants who were not given a loan.

Required

Using regression analysis to classify the potential borrowers, choose among several potential variables that might explain which variables **Lending Club** might use to decide whether or not to extend a loan.

Specify the Question

How can we use regression analysis to understand who receives a loan and who does not?

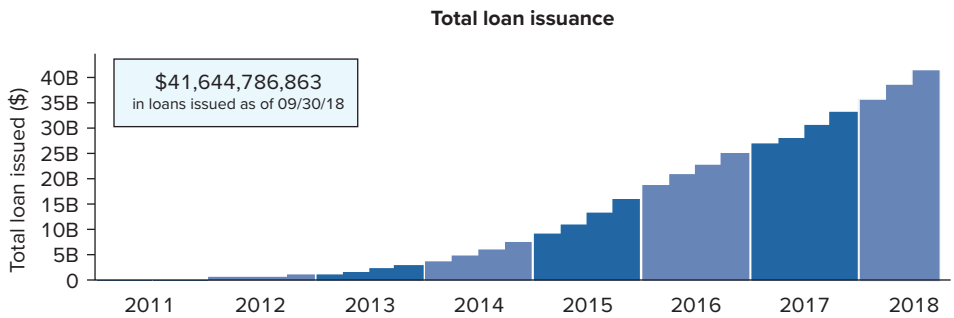
Obtain the Data

Lending Club is a U.S.-based peer-to-peer lending company, headquartered in San Francisco, California. Lending Club facilitates both borrowing and lending by providing a platform for unsecured personal loans between $1,000 and $35,000. The loan period is either three or five years. There is information available that allows potential investors to search and browse the loan listings on Lending Club's website and select loans in which they would like to invest. The website includes information about the borrower, amount of the loan, loan grade (and related loan interest rate), and loan purpose. Investors invest in the loans and make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee. Since 2007, hundreds of thousands of borrowers have obtained over $42 billion in loans via Lending Club.[9]

The basic statistics shown in Lab Exhibit 9.3.1 are included on the "Lending Club Statistics" web page.

The data set contains details on 645,414 loans (from 2007 to 2012) that were rejected and 39,708 loans (from 2007 to 2011) that were accepted.

Open Excel file Lab 9.3 Data.xlsx.



Lab Exhibit 9.3.1 Lending Club Statistics

(Source: https://www.lendingclub.com/info/statistics.action. Accessed 1/15/2019.)

[9]https://www.lendingclub.com/. Accessed 1/15/2019.

Data Dictionary

Accept:	1 If Lending Club approved loan; 0 if Lending Club rejected loan
loan_amnt:	The amount of the requested loan
DTI Bucket:	=3 if debt-to-annual income ratio is greater than 0.2
	=2 if debt-to-annual income ratio is greater than 0.1 but less than or equal to 0.2
	=1 if debt-to-annual income ratio is less than or equal to 0.1
Emp Length:	Length of employment in years (if reported less than 1, then equals 0.5; if 10+ years, then equals 10)

Analyze the Data

Analysis Task 1: Hypothesize the Relationships Between Loan Acceptance and Possible Determinants.

Although we can use several statistical methods to help classify the loan applications that Lending Club receives, we choose to use regression analysis.

Regression analysis uses mathematics and statistics to determine which of several variables have an impact on an outcome and the size of that impact. Regression analysis helps us answer the questions "Which factors matter most in predicting the outcome?" and "Which factors can we ignore?"

In this case, we are trying to figure out which factors lead to an accepted or rejected loan. We can think about this question in terms of an algebraic equation where y is the dependent (outcome) variable and x represents the independent variables, and where $y = f(x)$. Here, the dependent variable (y) is Acceptance and the independent variables (x) are three potential determinants of loan acceptance. In other words, we will run a regression where y, or Loan Accept = f(factors potentially predicting acceptance).

It is useful to think through the variables. For example, it seems that the greater the borrower's annual income, the more likely the borrower will be able to repay the loan.

- If the prospective borrower requests a larger loan, is Lending Club more likely or less likely to grant the loan? We expect that the bigger the loan, the less likely the loan will be paid back (and the less likely the loan will be accepted). Therefore, we expect a negative (−) relationship between Accept and loan amount (loan_amnt).
- If the prospective borrower has been employed by the same employer for a longer time, is Lending Club more likely or less likely to grant the loan? We expect that the longer the employment length, the more likely that Lending Club will grant the loan due to the borrower's job stability. Therefore, we expect a positive (+) relationship between Accept and employment length (Emp Length).
- If the prospective borrower has a high debt-to-annual income ratio (DTI Bucket), is Lending Club more likely or less likely to grant the loan? We expect that the larger the borrower's debt relative to annual income, the less likely that the loan will be paid back and the less likely Lending Club will grant the loan. Therefore, we expect a negative (−) relationship between Accept and DTI Bucket.

We summarize these relationships as follows:

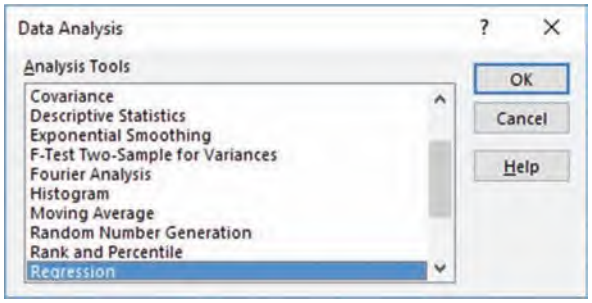
Independent variable	Expected relationship with Loan Acceptance (Accept)
Loan_amnt	–
Emp Length	+
DTI Bucket	–

Using regression analysis, we can assess which variables impact the Accept/Reject loan decision, and the size of that impact, by looking at the independent variables that are statistically related to the Accept decision.

Analysis Task 2: Perform the Regression.

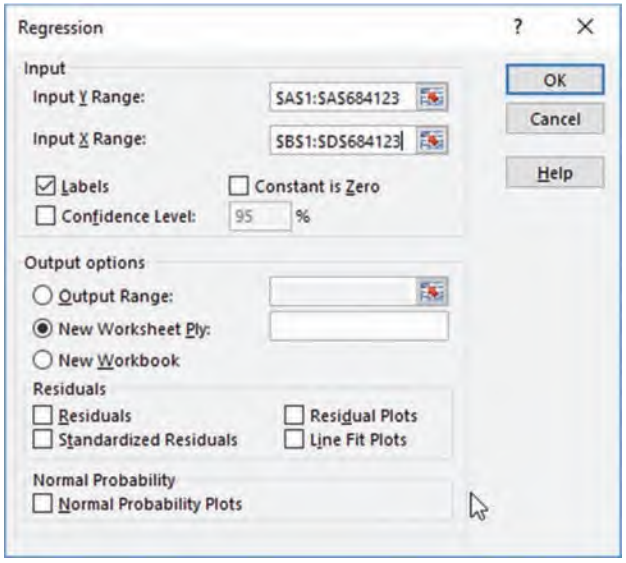
1. To run the descriptive statistics, we need to make sure the Data Analysis ToolPak is loaded. Look at the **Data > Analysis** and see if the Data Analysis add-in has been installed. If you haven't installed the Data Analysis ToolPak, textbook Appendix E provides instructions for installing it.
2. To perform the regression, select **Data> Analysis > Data Analysis**. A dialog box will open. Select Regression as shown in Lab Exhibit 9.3.2.

Lab Exhibit 9.3.2



3. A dialog box will open. We need to select the dependent variable (or Y Range). Our dependent (outcome) variable is Accept, so enter "$A$1:$A$684123" in the Input Y Range.
4. Next, we need to select our independent variable (or X Range). The variables loan_amnt, Emp Length, and DTI Bucket are our independent variables, so enter "$B$1:$D$684123" in the Input X Range.
5. Because we highlighted the labels in the data range, click on Labels. Also select New Worksheet Ply to get the results in a new worksheet.
6. The completed Regression dialog box should look as shown in Lab Exhibit 9.3.3.

Lab Exhibit 9.3.3



7. Now click OK to get the regression results shown in Lab Exhibit 9.3.4.

Lab Exhibit 9.3.4

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.303755
R Square	0.092267
Adjusted R Square	0.092263
Standard Error	0.220123
Observations	684122

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	3369.398	1123.133	23179.2	0
Residual	684118	33148.48	0.048454		
Total	684121	36517.88			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	0.07561	0.000815	92.72652	0	0.074011875	0.077208	0.074012	0.077208
loan_amnt	-9.1E-07	2.56E-08	-35.635	7.3E-278	-9.6382E-07	-8.6E-07	-9.6E-07	-8.6E-07
Emp Length	0.023968	9.7E-05	247.1157	0	0.023778042	0.024158	0.023778	0.024158
DTI Bucket	-0.02278	0.000322	-70.8158	0	-0.023407942	-0.02215	-0.02341	-0.02215

There are many things to note about the regression. The first is that the regression did better than chance at predicting loan acceptance. We draw this conclusion because the "Significance F" result is very small—in this case, zero—suggesting virtually zero probability that the loan acceptance/rejection decision (Accept) can be explained better by no independent variables than by a model that has independent variables. This result suggests that we will be able to identify the factors that explain loan acceptance.

We can use another statistic, Adjusted R Square, to measure how well the model predicted loan acceptance. *Note:* Statisticians prefer the term "adjusted R-squared," but Excel uses the term "Adjusted R Square" (see Lab Exhibit 9.3.4). The adjusted R-squared is a value between 0 and 1. An adjusted R-squared value of 0 indicates that the regression model has no ability to explain the dependent (outcome) variable and an adjusted R-squared value of 1 (or 100 percent) indicates that the model has perfect ability to explain the dependent variable. In this case, the adjusted R-squared value is 0.0923, which represents a relatively low ability to explain loan acceptance.

Although the overall explanatory power is low, we are still able to identify some predictors of loan acceptance. Specifically, the statistics regarding all three independent variables (loan_amnt, Emp Length, and DTI Bucket) suggest that they each help predict the acceptance decision. All have a t Stat that is greater than two (or less than negative two), suggesting that all three factors help predict the acceptance decision. As expected, the bigger the loan, the less likely the loan is accepted, as evidenced by the negative coefficient on the t Stat for loan_amnt. In addition, the longer the employment length, the more likely the loan will be accepted, as shown by the positive coefficient on Emp Length. Finally, as expected, the bigger the debt relative to income, the less likely the loan will be accepted, as shown in the negative coefficient on DTI Bucket.

Report the Results

We have now found three predictive factors for loan acceptance. Lending Club could use this information to help it decide whether to grant/reject a loan for a potential borrower.

Assessment

1. Take a screenshot of the regression output for all three independent variables, paste it into a Word document named "Lab 9.3 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Lab 9.3 Excel Submission.docx via Connect, if assigned.

There is no alternate Lab 9.3 Excel.

Lab 9.4 Excel

Predictive Analytics: Predicting Bankruptcy

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Predicting Bankruptcy, Distress, Classification

Decision-Making Context

In 1968, Edward Altman published a paper[10] with a model that predicts bankruptcy using an analysis called Altman's Z. Today, more than 50 years later, the basic bankruptcy prediction model is still used.

Analysis with Altman's Z is a classification exercise because it puts companies into classes, specifically the two classes of companies that will go bankrupt and companies that will not.

Altman's original research found that a combination of five common business ratios predicts bankruptcy:

1. X_1: Working capital ÷ Total assets
   - Measures liquid, cash-like assets (or liquidity level in relation to the size of the company)
2. X_2: Retained earnings ÷ Total assets
   - Measures the company's long-term profitability
3. X_3: Earnings before interest and taxes ÷ Total assets
   - Measures the company's recent, or short-term, profitability
4. X_4: Market value of stock ÷ Book value of liabilities
   - Measures the company's long-term solvency, or whether the company will have sufficient funds to pay its debt as it comes due
5. X_5: Sales ÷ Total assets
   - Measures efficiency, or how well assets are utilized

Required

1. Compute Altman's five factors used to predict bankruptcy.
2. Weight each of those factors using Altman's bankruptcy prediction weights to arrive at Altman's Z score.
3. Use a histogram to classify each company as either in the distress zone, gray zone, or safe zone for bankruptcy.

Specify the Question

Which companies do we predict will go bankrupt?

Obtain the Data

Financial data are provided for publicly traded companies in the retail sector. Open Excel file Lab 9.4 Data.xlsx and view the variables. The data set contains details on 5,437 companies with all necessary data for the years from 2017 to 2018. Lab Exhibit 9.4.1 provides a snippet of the data.

[10]Altman, Edward I., "Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy," *The Journal of Finance,* 23.4 (1968), 589–609.

	A	B	C	D	E	F	G	H	I	J	K	L
1	gvkey	conm	fyear	act	at	ebit	lct	lt	ni	re	sale	ME
2	1004	AAR CORP	2017	942.7	1524.7	86	333.3	588.4	15.6	701.2	1748.3	1551.458
3	1045	AMERICAN AIRLINES GROUP INC	2018	8637	60580	3449	18096	60749	1412	-5138	44541	14790.22
4	1050	CECO ENVIRONMENTAL CORP	2018	160.865	392.582	14.062	103.886	214.022	-7.121	-72.842	337.339	235.008
5	1072	AVX CORP	2017	1701.724	2672.766	182.534	294.563	429.323	4.91	1983.724	1562.474	2787.583
6	1075	PINNACLE WEST CAPITAL CORP	2018	924.991	17664.202	773.687	1648.964	12315.497	511.047	2593.475	3691.247	9551.09
7	1078	ABBOTT LABORATORIES	2018	14632	67173	4284	9012	36451	2368	16974	30578	126983.9
8	1084	WORLDS INC	2018	3.846	3.846	-2.893	4.441	4.441	1.754	-41.165	0	11.93115
9	1094	ACETO CORP	2018	502.627	767.024	-17.187	302.518	671.739	-316.121	-127.622	711.359	103.1365
10	1096	MORGUARD CORP	2018	204.517	11082.758	451.277	1321.718	6436.479	319.851	3328.93	1157.878	1987.857
11	1104	ACME UNITED CORP	2018	72.992	109.477	7.456	13.365	57.145	4.598	45.492	137.321	47.75175
12	1117	BK TECHNOLOGIES CORP	2018	30.856	39.191	2.425	9.883	11.479	-0.195	-2.393	49.38	48.0675
13	1121	ADAMS RESOURCES & ENERGY INC	2018	229.59	278.87	0.798	123.267	132.272	2.945	134.228	1750.213	163.2788
14	1161	ADVANCED MICRO DEVICES	2018	3540	4556	496	1984	3290	337	-7444	6475	18552.3
15	1166	ASM INTERNATIONAL NV	2018	779.245	2115.739	142.477	222.398	236.331	179.901	2194.09	936.619	2040.919
16	1186	AGNICO EAGLE MINES LTD	2018	1072.452	7852.843	92.17	361.449	3302.831	-326.701	-1047.008	2191.221	9472.144
17	1209	AIR PRODUCTS & CHEMICALS INC	2018	5082.2	19178.3	1891.3	2338.3	8002	1497.8	11668	8930.2	36670.15

Lab Exhibit 9.4.1

Data Dictionary

gvkey: A unique code for each company given by Compustat, the data provider for these financial statement data

conm: Company name

fyear = The fiscal year

act: Current assets ($ millions)

at: Total assets ($ millions)

ebit: Earnings before interest and taxes ($ millions)

lct: Current liabilities ($ millions)

lt: Total liabilities ($ millions)

ni: Net income ($ millions)

re: Retained earnings ($ millions)

sale: Net sales ($ millions)

ME: Market value of equity ($ millions)

Analyze the Data

Analysis Task 1: Create Each of the Five Factors.

1. In cell M1, insert the title "X1".
2. In cell N1, insert the title "X2".
3. In cell O1, insert the title "X3".
4. In cell P1, insert the title "X4".
5. In cell Q1, insert the title "X5".
6. In cell R1, insert the title "Altman's Z".
7. In Excel, calculate the five financial ratios (labeled as X1:X5) useful in predicting bankruptcy as follows:
   X1: Working capital / Total assets: Calculate in cell M2 using the variable names in the data dictionary as "=(act-lct)/at".
   - Measures liquid, cash-like assets (or liquidity level in relation to the size of the company)

   X2: Retained earnings / Total assets: Calculate in cell N2 using the variable names in the data dictionary as "=re/at".
   - Measures the company's long-term profitability

X3: Earnings before interest and taxes / Total assets: Calculate in cell O2 using the variable names in the data dictionary as "=ebit/at".

- Measures the company's recent, or short-term, profitability

X4: Market value of stock / Book value of liabilities: Calculate in cell P2 using the variable names in the data dictionary as "=ME/lt".

- Measures the company's long-term solvency, or whether the company will have sufficient funds to pay its debt as it comes due

X5: Sales / Total assets: Calculate in cell Q2 using the variable names in the data dictionary as "=sale/at".

- Measures efficiency, or how well assets are utilized

The formulas in the Excel spreadsheet should look as shown in Lab Exhibit 9.4.2.

Lab Exhibit 9.4.2

	M	N	O	P	Q
1	X1	X2	X3	X4	X5
2	=(D2-G2)/E2	=J2/E2	=F2/E2	=L2/H2	=K2/E2

Copy the results from L1:Q1 to L2:Q5438 to get the result shown in Lab Exhibit 9.4.3.

Lab Exhibit 9.4.3

	M	N	O	P	Q	R
1	X1	X2	X3	X4	X5	ALTMAN'S
2	$0.40	$0.46	$0.06	2.63674	1.146652	
3	($0.16)	($0.08)	$0.06	0.243464	0.735243	
4	$0.15	($0.19)	$0.04	1.098055	0.859283	
5	$0.53	$0.74	$0.07	6.492973	0.584591	
6	($0.04)	$0.15	$0.04	0.775534	0.208968	
7	$0.08	$0.25	$0.06	3.483688	0.455213	
8	($0.15)	($10.70)	($0.75)	2.686591	0	
9	$0.26	($0.17)	($0.02)	0.153536	0.927427	
10	($0.10)	$0.30	$0.04	0.308842	0.104476	

Analysis Task 2: Calculate Altman's Z to Predict the Chance of Bankruptcy.

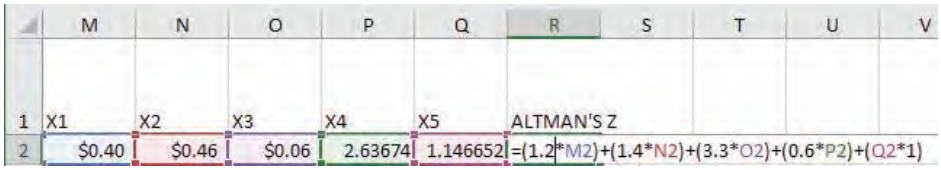
Altman's Z is a bankruptcy prediction using a combination of five common business ratios. In essence, the higher each of these common business ratios, the less likely the company will go bankrupt.

The original Z-score formula is:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5.$$

1. In cell R2, calculate the Z-score by weighting each of the factors X1–X5, as shown in Lab Exhibit 9.4.4.

Lab Exhibit 9.4.4



2. Copy the results in cell R2 to cells R3:R5438 to get the result shown in Lab Exhibit 9.4.5.

Lab Exhibit 9.4.5

	M	N	O	P	Q	R
1	X1	X2	X3	X4	X5	ALTMAN'S
2	$0.40	$0.46	$0.06	2.63674	1.146652	4.038304
3	($0.16)	($0.08)	$0.06	0.243464	0.735243	0.763093
4	$0.15	($0.19)	$0.04	1.098055	0.859283	1.550722
5	$0.53	$0.74	$0.07	6.492973	0.584591	6.376601
6	($0.04)	$0.15	$0.04	0.775534	0.208968	0.975194
7	$0.08	$0.25	$0.06	3.483688	0.455213	3.21005
8	($0.15)	($10.70)	($0.75)	2.686591	0	-16.0406
9	$0.26	($0.17)	($0.02)	0.153536	0.927427	1.025733
10	($0.10)	$0.30	$0.04	0.308842	0.104476	0.723705

Altman found he could accurately classify the companies into three zones, or classes, using the following decision rules:

If $Z < 1.80$ ⇒ Classify as significant risk of bankruptcy, the "distress zone"
If $Z \geq 1.80$ and $Z < 3.0$ ⇒ Classify as at risk of bankruptcy, the "gray zone"
If $Z \geq 3.00$ ⇒ Classify as not currently at risk of bankruptcy, the "safe zone"

To evaluate the Altman Z-scores to see if any fall in these zones, let's do additional analysis that includes a histogram to count how many are in each bin, or zone.



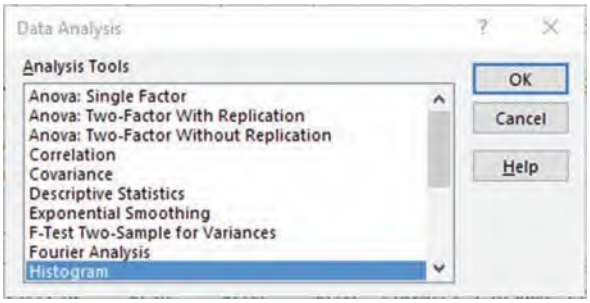
DATA VISUALIZATION

Analysis Task 3: Create a Histogram to Evaluate the Frequency of Firms in Each Bankruptcy Zone.

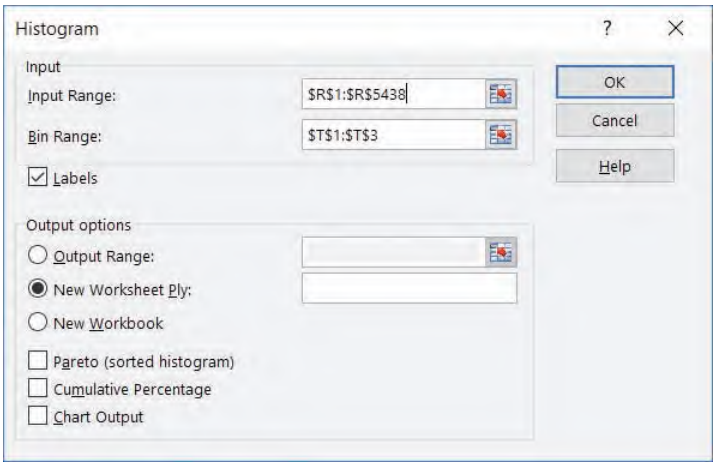
To create the histogram, we make sure the Data Analysis ToolPak is loaded. Look at the **Data > Analysis** and see if the Data Analysis add-in has been installed. If the Data Analysis ToolPak isn't installed, textbook Appendix E provides instructions for installing it.

1. In cell T1, insert the title “Bins”.
2. In cell T2, insert the number “1.799999”. This will be the highest number of the lowest bin to help count the number of companies that fall in the Distress Zone of Altman’s Z.
3. In cell T3, insert the number “2.999999”. This will be the highest number of the middle bin to help count the number of companies that fall in the Gray Zone of Altman’s Z.
4. We’ll now create a histogram. Select **Data > Analysis > Data Analysis** and select Histogram. Then click OK (Lab Exhibit 9.4.6).
5. Input “$R$1:$R$5438” in the Input Range and “$T$1:$T$3” in the Bin Range and select Labels. These instructions tell the histogram where to look for the data and what bins it should use for the various zones included in the data. Also select “New Worksheet Ply” to get the results in a new worksheet. The Histogram dialog box should appear as shown in Lab Exhibit 9.4.7. Click OK.

Lab Exhibit 9.4.6



Lab Exhibit 9.4.7



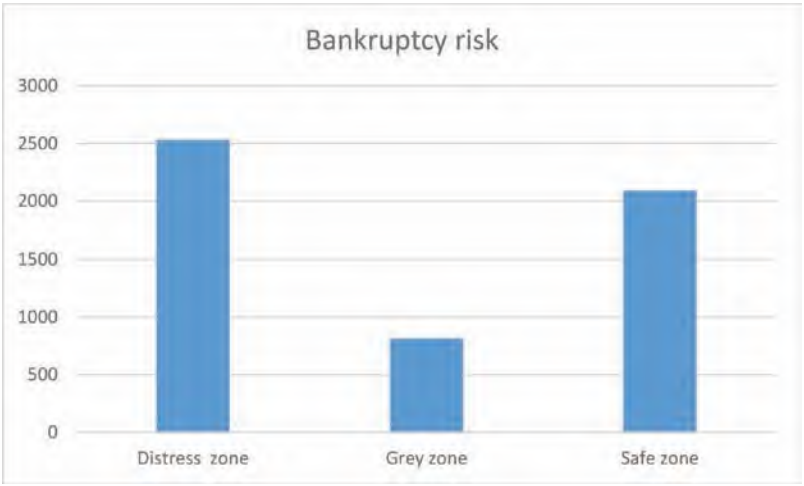
6. The resulting table histogram should show that there are 2,530 firms in the distress zone, 817 firms in the gray zone, and 2,090 firms in the safe zone, as summarized in Lab Exhibit 9.4.8.

Lab Exhibit 9.4.8

	A	B
1	*Bins*	*Frequency*
2	1.799999	2530
3	2.999999	817
4	More	2090

7. Rename the bins as “Distress zone”, “Grey zone”, and “Safe zone”. Insert a bar chart and retitle the bar chart as “Bankruptcy risk”, as shown in Lab Exhibit 9.4.9.

Lab Exhibit 9.4.9



Report the Results

We now have a bankruptcy prediction for every retail company. Auditors can use this information to understand the financial risk facing their clients. Investors can use this information to understand the financial risk they face if they invest in the firm. Banks and lenders can also use this information to decide if the company will be around to pay the loan back when it is due.

Assessment

1. Take a screenshot of the top 10 rows of calculations in your spreadsheet (including columns including X1:X5 and the Altman Z calculation), paste it into a Word document named “Lab 9.4 Excel Submission.docx”, and label the screenshot Submission 1.
2. Take a screenshot of the histogram showing the distribution of Altman’s Z scores for this data set, paste it into the same document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Lab 9.4 Excel Submission.docx via Connect, if assigned.

Alternate Lab 9.4 Excel

Predictive Analytics: Predicting Bankruptcy

Following the instructions in Lab 9.4, analyze the chance of bankruptcy for manufacturing firms included in the Alt Lab 9.4 Data.xlsx data set.

Required

1. Compute Altman’s five factors used to predict bankruptcy.
2. Weight each of those factors using Altman’s bankruptcy prediction weights to arrive at Altman’s Z score.
3. Use a histogram to classify each company as either in the distress zone, gray zone, or safe zone for bankruptcy.

Open Excel file Alt Lab 9.4 Data.xlsx.

Assessment

1. Take a screenshot of the top 10 rows of calculations in your spreadsheet (including columns including X1:X5 and the Altman Z calculation), paste it into a Word document named “Alt Lab 9.4 Excel Submission.docx”, and label the screenshot Submission 1.
2. Take a screenshot of the histogram showing the distribution of Altman’s Z-scores for this data set, paste it into the same document, and label the screenshot Submission 2.
3. Answer the questions in Connect and upload your Alt Lab 9.4 Excel Submission.docx via Connect, if assigned.

Lab 9.5 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Prescriptive Analytics: Evaluating Investments Using NPV

Keywords

NPV(), Discounting Cash Flows, Investment Evaluation

Decision-Making Context

Managers can use a few tools to evaluate the profitability of potential investments. The net present value (or NPV) is calculated as the present value of cash inflows minus the present value of the cash outflows. We use the company's cost of capital as the benchmark for the cost of acquiring the investment. In general, if the net present value is not greater than the costs of acquiring the investment, there is no reason to pursue that investment.

Net present value also helps decision-makers rank various alternate, competing investments. Although investors can't predict the future with certainty, they can use all of the information they have at their disposal, including the net present value, to evaluate potential investments.

In this lab, we take the perspective of a manager looking at two different potential investments in shares of stock and considering the profitability of the expected cash flows.

Required

Use NPV to evaluate the potential investments.

Specify the Question

How can we use NPV to evaluate potential investments?

Obtain the Data

Open Excel file Lab 9.5 Data.xlsx.

Suppose there are four potential stock investments, with each one costing $50,000 upfront. The investor has a five-year horizon, meaning he plans to sell the investment after five years. The first, third, and fourth companies pay dividends, but the second company currently does not.

Dividends: A sum of money paid regularly (typically quarterly) by a company to its shareholders out of its profits.

The analysis suggests the projected string of cash flows shown in Lab Exhibit 9.5.1.

Lab Exhibit 9.5.1

	A	B	C	D	E	F
1		Cash Flow	Investment	Investment	Investment	Investment
2	Year	Description	1	2	3	4
3	Year 0 (today)	Buy Stock	(50,000)	(50,000)	(50,000)	(50,000)
4	Year 1	Dividends	10,000	-	-	15,000
5	Year 2	Dividends	10,000	-	-	-
6	Year 3	Dividends	10,000	-	-	-
7	Year 4	Dividends	10,000	-	15,000	-
8	Year 5	Sell Stock	100,000	139,000	126,000	128,000

Note that Year 0 is today, or the date of the stock purchase. Years 1-4 are years when dividends are paid. Year 5 is the year that the stock is sold.

Also note that the $50,000 investment is shown as a negative number because cash is being paid out instead of being received.

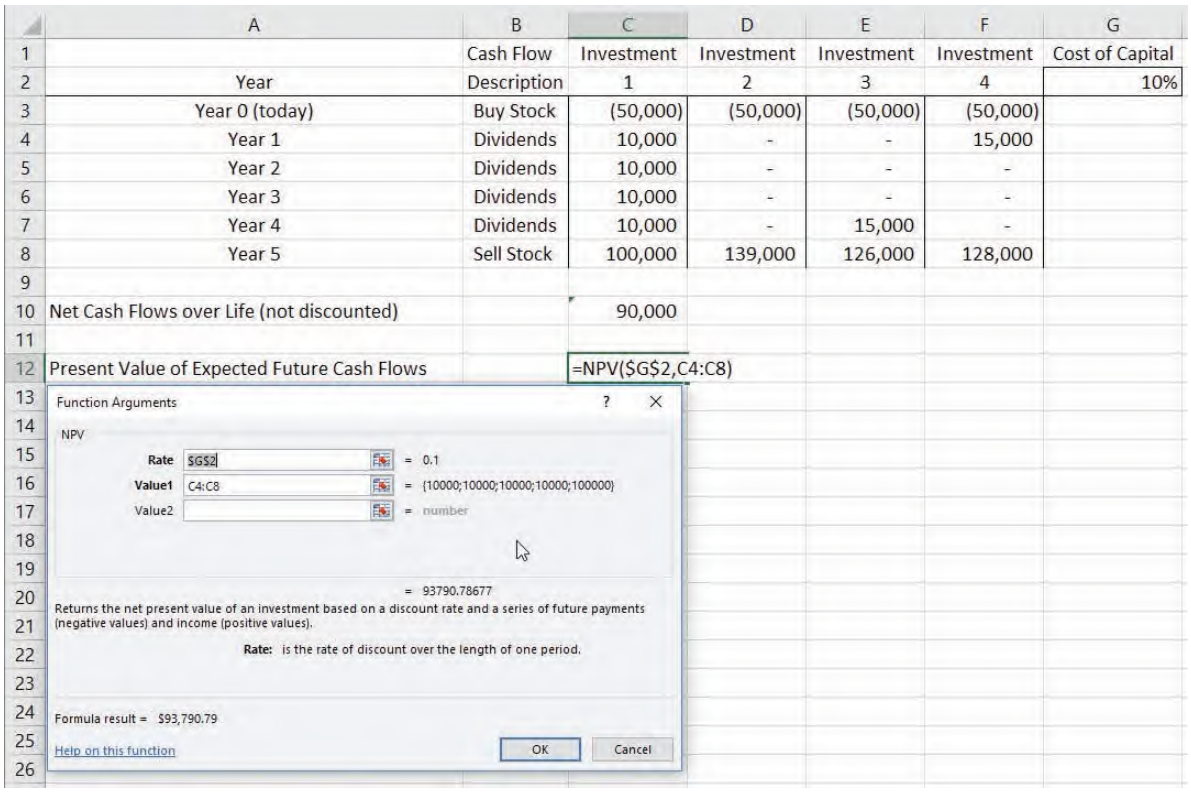
We can use the net present value (NPV) to compare the cash flows of potential investments by evaluating the amounts and timing of those cash flows. Because the amount of cash returned under each investment scenario is the same in this case, the timing of those cash inflows is particularly important.

We assume the cost of capital in this case to be 10 percent.

Analyze the Data

Analysis Task 1: Calculate Net Present Value for the Four Potential Investments.

1. In cell A10, insert the title "Net Cash Flows over Life (not discounted)".
2. In cell C10, sum the cash flows from C3:C8. Copy the contents of cell C10 to cells D10:F10.
3. In cell G1, insert the title "Cost of Capital".
4. In cell G2, insert the number "10%", which is the cost of capital for this company.
5. In cell A12, insert the title "Present Value of Expected Future Cash Flows".
   There is an Excel function called NPV() that we can use to evaluate the return for these investments.
   The =NPV() function requires information regarding the rate used to denote the cost of acquiring the funds or the cost of capital as well as the string of cash flows.
6. In cell C12, insert "=NPV($G$2,C4:C8)", where $G$2 refers to the cost of capital and C4:C8 refers to the string of cash flows. Note that we use an absolute reference to $G$2, so that it will stay referenced to the cost of capital as we copy to other cells (Lab Exhibit 9.5.2).



Lab Exhibit 9.5.2

7. We use the same NPV function for Investments 2–4 by copying the contents of C12 to cells D12:F12.
8. In cell A13, insert the title "Initial Outlay".
9. In cell C13, insert the equation "=C3".
10. In cell A14, insert the title "Net Present Value (NPV) of Investment".
11. In cell C13, insert the equation "=SUM(C12:C13)".
12. Copy the result from C13:C14 to D13:F14.

Your spreadsheet should now look as shown in Lab Exhibit 9.5.3.

Lab Exhibit 9.5.3

	A	B	C	D	E	F	G
1		Cash Flow	Investment	Investment	Investment	Investment	Cost of Capital
2	Year	Description	1	2	3	4	10%
3	Year 0 (today)	Buy Stock	(50,000)	(50,000)	(50,000)	(50,000)	
4	Year 1	Dividends	10,000	-	-	15,000	
5	Year 2	Dividends	10,000	-	-	-	
6	Year 3	Dividends	10,000	-	-	-	
7	Year 4	Dividends	10,000	-	15,000	-	
8	Year 5	Sell Stock	100,000	139,000	126,000	128,000	
9							
10	Net Cash Flows over Life (not discounted)		90,000	89,000	91,000	93,000	
11							
12	Present Value of Expected Future Cash Flows		$93,790.79	$86,308.06	$88,481.29	$93,114.29	
13	Initial Outlay		(50,000)	(50,000)	(50,000)	(50,000)	
14	Net Present Value (NPV) of Investment		$43,790.79	$36,308.06	$38,481.29	$43,114.29	

We find that the net present value is higher for Investment 1 at $43,790.79 than it is for all the other investments. Why? The time value of money suggests that the sooner we get the money, the sooner we can reinvest it in other investments or use it for other purposes.

Report the Results

We have done a net present value analysis to evaluate the profitability of Investments 1–4. What we have not factored in here is the different amount of risk for these four companies. Likely the best way to present this analysis to decision-makers is to show a table of the expected cash flows, along with the resulting net present values achieved in each scenario.

Different investors have different views. For example, some like receiving dividends as they go (as in Investments 1, 3, and 4) and view that situation as less risky than waiting to get the entire investment returned at the end (as in Investment 2). However, not knowing more about the companies, it is hard to understand the level of risk involved. For example, if Investment 1 is much riskier than Investment 4, then perhaps Investment 4 would be a better investment. If the levels of risk differ, then we can incorporate that difference into the cost of acquiring the funds. We can use other metrics (such as internal rate of return and risk-adjusted returns) to help incorporate risk in our analysis.

Assessment

1. Take a screenshot of the NPV analysis, paste it into a Word document named "Lab 9.5 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Lab 9.5 Excel Submission.docx via Connect, if assigned.

Alternate Lab 9.5 Excel

Prescriptive Analytics: Evaluating Investments Using NPV

Apply the same steps in Lab 9.5 Excel to the Alt Lab 9.5 Data.xlsx data set.

Vern Richardson is trying to evaluate the profitability of four investments over an eight-year horizon. Use the Excel NPV() function to evaluate these four potential investments over the life of their cash flows.

Required

Use NPV to evaluate the potential investments.

Open Excel file Alt Lab 9.5 Data.xlsx and perform the NPV analysis. Assume the cost of capital is 6 percent.

Assessment

1. Take a screenshot of the NPV analysis, paste it into a Word document named "Alt Lab 9.5 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 9.5 Excel Submission.docx via Connect, if assigned.

Lab 9.6 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Prescriptive Analytics: Evaluating Investments Using IRR

Keywords

IRR(), Discounting Cash Flows, Investment Evaluation

Decision-Making Context

Investors can use a few tools to evaluate the profitability of potential investments. The internal rate of return (or IRR) is a discount rate that makes the net present value of all cash flows from an investment or project equal to zero.

Although investors can't predict the future with certainty, they use all of the information they have at their disposal, including the internal rate of return, to evaluate potential investments.

In this lab, we take the perspective of an investor looking at four different potential investments in shares of stock and considering the profitability of the expected cash flows.

Required

Use IRR to evaluate the potential investments.

Specify the Question

How can we use IRR to evaluate potential investments?

Obtain the Data

Open Excel file Lab 9.6 Data.xlsx.

Suppose there are four potential investments, with each one costing $50,000 up front. The investor has a five-year horizon, meaning she plans to sell the investment after five years. The first, third, and fourth companies pay dividends, but the second company currently does not.

The analysis suggests the projected string of cash flows in Lab Exhibit 9.6.1.

Lab Exhibit 9.6.1

	A	B	C	D	E	F
1		Cash Flow	Investment	Investment	Investment	Investment
2	Year	Description	1	2	3	4
3	Year 0 (today)	Buy Stock	(50,000)	(50,000)	(50,000)	(50,000)
4	Year 1	Dividends	10,000	-	-	15,000
5	Year 2	Dividends	10,000	-	-	-
6	Year 3	Dividends	10,000	-	-	-
7	Year 4	Dividends	10,000	-	15,000	-
8	Year 5	Sell Stock	100,000	139,000	126,000	128,000

Note that Year 0 is the date of the investment. Years 1–4 are years when dividends are paid. Year 5 is the year that the stock is sold.

Also note that the $50,000 investment is shown as a negative number because cash is being paid out (as a cash outflow) instead of cash received.

We can use the internal rate of return (sometimes called IRR) to compare the cash flows of potential investments by evaluating the amounts and timing of those cash flows.

Analyze the Data

Analysis Task 1: Calculate the Internal Rate of Return for the Four Potential Investments.

There is an Excel function called IRR() that we can use to evaluate the return for these investments (Lab Exhibit 9.6.2, Row 12).

Lab Exhibit 9.6.2

	A	B	C	D	E	F
1		Cash Flow	Investment	Investment	Investment	Investment
2	Year	Description	1	2	3	4
3	Year 0 (today)	Buy Stock	(50,000)	(50,000)	(50,000)	(50,000)
4	Year 1	Dividends	10,000	-	-	15,000
5	Year 2	Dividends	10,000	-	-	-
6	Year 3	Dividends	10,000	-	-	-
7	Year 4	Dividends	10,000	-	15,000	-
8	Year 5	Sell Stock	100,000	139,000	126,000	128,000
9						
10	Net Cash Flows over Life (not discounted)		90,000	89,000	91,000	93,000
11						
12	Internal Rate of Return (IRR)		=IRR(C3:C8)			

1. In cell A10, insert the words "Net Cash Flows over Life (not discounted)".
2. In cell C10, sum the cash flows from C3:C8. Copy the contents of cell C10 to cells D10:F10.
3. In cell A12, insert the words "Internal Rate of Return (IRR)".
4. In cell C12, insert "=IRR(C3:C8)".
5. We use the same IRR function for Investments 2, 3, and 4 by copying the contents of cell C12 to cells D12:F12.

Your spreadsheet should now look as shown in Lab Exhibit 9.6.3.

Lab Exhibit 9.6.3

	A	B	C	D	E	F
1		Cash Flow	Investment	Investment	Investment	Investment
2	Year	Description	1	2	3	4
3	Year 0 (today)	Buy Stock	(50,000)	(50,000)	(50,000)	(50,000)
4	Year 1	Dividends	10,000	-	-	15,000
5	Year 2	Dividends	10,000	-	-	-
6	Year 3	Dividends	10,000	-	-	-
7	Year 4	Dividends	10,000	-	15,000	-
8	Year 5	Sell Stock	100,000	139,000	126,000	128,000
9						
10	Net Cash Flows over Life (not discounted)		90,000	89,000	91,000	93,000
11						
12	Internal Rate of Return (IRR)		29.0%	22.7%	23.7%	27.3%

We find that the internal rate of return is higher for Investment 1 (29 percent) than it is for the other investments. Why? First, the amount of the cash flows, and second, the timing of those cash flows arriving. Indeed, the time value of money suggests that the sooner we get the money, the sooner we can reinvest it in other investments or use it for other purposes.

Report the Results

We have done one type of analysis—calculating the internal rate of return—to evaluate the profitability of Investments 1–4. What we have not factored in here is the different amount of risk for these four companies.

Different investors have different views. For example, some like receiving dividends as they go (as in Investments 1, 3, and 4) and view that situation as less risky than waiting to get the entire investment returned at the end (as in Investment 2). However, not knowing more about the companies, it is hard to understand the level of risk involved. For example, if Investment 1 is much riskier than Investment 4, then perhaps Investment 4 would be a better investment. If the levels of risk differ, then we can incorporate that difference into the cost of acquiring the funds. We can use other metrics (such as net present value and risk-adjusted returns) to help incorporate risk in our analysis.

Assessment

1. Take a screenshot of the IRR analysis, paste it into a Word document named "Lab 9.6 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Lab 9.6 Excel Submission.docx via Connect, if assigned.

Alternate Lab 9.6 Excel

Prescriptive Analytics: Evaluating Investments Using IRR

Apply the same steps in Lab 9.6 to the Alt Lab 9.6 Data.xlsx data set.

Vern Richardson is trying to evaluate the profitability of four investments over an eight-year horizon. Use the Excel IRR() function to evaluate these four potential investments over the life of their cash flows.

Required

Use IRR to evaluate the potential investments.

Open Excel file Alt Lab 9.6 Data.xlsx and perform the IRR analysis.

Assessment

1. Take a screenshot of the IRR analysis, paste it into a Word document named "Alt Lab 9.6 Excel Submission.docx", and label the screenshot Submission 1.
2. Answer the questions in Connect and upload your Alt Lab 9.6 Excel Submission.docx via Connect, if assigned.

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Chapter 10

Operations Analytics

A Look Back

- Chapters 7 through 9 described how companies use business analytics to gain actionable business intelligence (BI) across three main business functions: marketing, accounting, and finance. Each chapter showed how business analysts can use the SOAR model to ask and answer questions.
- Chapter 9 discussed examples of finance questions addressable using business analytics, the sources of financial data, the key types of financial analytics, and how the results of financial analytics can be reported.

A Look at This Chapter

- Operations turn the right inputs into the right goods and services that companies sell. The SOAR model can help business analysts ask and answer questions in three important branches: human resources, information technology (IT) operations, and supply chain.
- In this textbook, we refer to these three areas collectively as *operations* because they provide the people, IT, and product production and distribution that the company needs in order to operate.

A Look Ahead

- Advances in technology are creating a fifth type of analytics, adaptive/autonomous analytics.
- Adaptive/autonomous analytics mimics human capabilities, such as judgment and reasoning, creating insights from structured and unstructured data. It takes analytics to a new level as reiterative technology executes decisions, scans the environment, adapts to changes, and then executes new decisions.
- Chapter 11 provides examples of business applications of advanced analytics, describing how companies can pull everything together using robotic process automation and cockpits. It also discusses the future of business analytics.

The Demise of Google's Human Resource Zingers

Can you imagine being asked the following questions during a job interview?

Question 1: "A book has N pages, numbered the usual way, from 1 to N. The total number of digits in the page numbers is 1,095. How many pages does the book have?"

Question 2: "A man pushed his car to a hotel and lost his fortune. What happened?"

At one time, answers to these questions affected whether or not **Google** hired a job candidate. However, thanks to human resource (HR) analytics, Google no longer asks these questions of its candidates. After analyzing data from tens of thousands of job interviews, Google realized that answering such brainteaser questions correctly did not predict long-term success in the company. Analytics also showed that college grade point averages (GPAs) and test scores were not good predictors of job performance either, unless the applicant had just been awarded a degree. The results from HR analytics caused Google to change its approach to interviews to a structured behavioral approach, which focuses on past performance and behavior with questions such as "Describe a time when you were under a lot of pressure at work. How did you react?" As this example shows, HR analytics can help transform companies into higher-performing organizations.[1]

googol = a very large number (10^{100})=

10000000000000000000000
00000000000000000000000
00000000000000000000000
00000000000000000000000
00000000000000000000000

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

LO 10.1 Explain operations and its three branches: human resources, information technology operations, and supply chain.

LO 10.2 Identify representative operations questions addressable by business analytics in each branch of operations.

LO 10.3 Identify internal and external sources of operations data.

LO 10.4 Describe and provide examples of descriptive analytics used to answer operations questions.

LO 10.5 Describe and provide examples of diagnostic analytics used to answer operations questions.

LO 10.6 Describe and provide examples of predictive analytics used to answer operations questions.

LO 10.7 Describe and provide examples of prescriptive analytics used to answer operations questions.

LO 10.8 Explain how to report the results of operations analytics.

LABS FOR CHAPTER 10

LAB 10.1 **Excel:** Descriptive Analytics/Supply Chain: Calculating KPIs for Walmart and Amazon (or Bed, Bath, and Beyond and Target)

LAB 10.2 **Tableau/Power BI:** Descriptive Analytics/Human Resources: Analyzing Employee Turnover Using a Bar Chart

LAB 10.3 **Excel:** Diagnostic Analytics/Supply Chain: Analyzing KPI Differences Between Walmart and Amazon (or Bed, Bath, and Beyond and Target)

LAB 10.4 **Tableau:** Diagnostic Analytics/IT Operations: Analyzing Website Response Time Using a Control Chart

LAB 10.5 **Tableau:** Predictive Analytics/Supply Chain: Forecasting Unit Demand

LAB 10.6 **Excel:** Prescriptive Analytics/Human Resources: Optimizing Employee Assignment

[1]In case you are wondering, the answers to the questions above are as follows: Answer 1: 401 pages. Answer 2: He was playing Monopoly.

SOARing to Success

Fun Fact: What Is a Googol?

According to some stories, Google was originally just a misspelled word. In 1997, Larry Page and Sergey Brin were looking for a name for their new search engine. They wanted a term suggesting the indexing of large amounts of data. Larry came up with googol, a very large number (10^{100}). Sean mistakenly typed in google.com and, finding that the website was available, registered their new search engine. And the rest is Google history![2]

LO 10.1

Explain operations and its three branches: human resources, information technology operations, and supply chain.

10.1 THE ROLE OF OPERATIONS IN BUSINESS

"People think we got big by putting big stores in small towns. Really, we got big by replacing inventory with information."

—SAM WALTON, FOUNDER OF **Walmart**

Operations are the processes that turn inputs into goods and services that companies sell. Operations ensures that an organization's materials, labor, and other inputs are used in the most effective and efficient way possible, thus maximizing a company's output.[3] In other words, operations is the backbone of business efficiency[4] and essential to the company's success.

How do companies use operations? Here are a few examples from each branch of operations:

Supply Chain:

- **Dell** uses operations to turn components into finished computers.
- **Home Depot** uses operations to buy inventory to sell in its stores and online.
- **Aerosoles.com** uses operations to ship products to customers.

Human Resources:

- **Exxon** uses operations to provide benefits to its employees.
- **Anheuser-Busch** uses operations to hire new employees.

IT Operations:

- **Forbes** uses operations to keep its websites up and running.
- **EY** uses operations to connect audit teams across the globe.

The Three Primary Branches of Operations

In this chapter, we examine three primary branches of operations: human resources, IT operations, and supply chain.

Human resources (HR) is the branch of operations that focuses on managing and developing a company's workforce. HR processes include recruiting, hiring, compensating, and training employees as well as managing their performance.

IT operations focuses on how the company manages its hardware, software, and IT support. IT operations is also responsible for collecting, storing, securing, and managing data. For example, an organization's policies about cloud computing, IoT (the Internet of

[2]Source: https://graphics.stanford.edu/~dk/google_name_origin.html. Accessed 10/07/2021.
[3]Source: William J. Stevenson, 2021 *Operations Management* (14th Edition). New York, NY: McGraw Hill.
[4]Source: Becky Leighton, May 9, 2020, "Operations Management: The Backbone of Business Efficiency." Available online at https://masterstart.com/blog/business/operations-management-the-backbone-of-business-efficiency/. Accessed 10/07/2021.

Things, which is composed of devices connected to the internet), and machine-to-machine communications are created by the managers in charge of IT operations. The branch in charge of a company's information technology, or IT, goes by many different names, including Integrated Technology Services, Global Technology Solutions, Digital Technology Solutions, and the Business Technology Unit. Companies may also split IT into operations, applications, and help desk groups. In this book, we refer to the department(s) in charge of anything related to IT as IT operations.

Supply chain is the branch of operations that is responsible for the production of a product or service and its distribution to the final consumer. It includes all of the processes necessary to make the product: obtaining raw materials, transforming those materials into the finished product, and transporting and distributing those products. An organization's supply chain is the network of entities that are involved in executing the process, including vendors, transportation companies, warehouses, distribution centers, and retailers. Supply chains are integral parts of most organizations' business and are essential in developing customer satisfaction and organizational success.[5] Without these networks it would be impossible to procure and distribute goods. Exhibit 10.1 shows the seven steps in the supply chain of a manufacturing organization.



Exhibit 10.1
Sample Supply Chain with Entities and the Seven Steps

Adapted from: William Vorhies. 2015. "Predictive Analytics in the Supply Chain." Available online at https://www.datasciencecentral.com/profiles/blogs/predictive-analytics-in-the-supply-chain. Accessed on 10/07/2021.

Retailers that buy finished goods rather than manufacturing their own products also have supply chains. However, their supply chain starts with purchasing finished goods instead of buying raw materials. Service organizations have modified supply chains as well. They typically rely on labor and human interactions rather than investments in plant and equipment to fill customers' or clients' demand for a service. For example, a marketing client may ask an advertising agency for a new advertising campaign. Based on the client's specifications, the agency will create a marketing plan, which may include a logo, website, and online advertisements. All of the steps to create the marketing plan are the service supply chain. In this textbook, we focus on a typical manufacturing supply chain.

[5]Source: James R. Stock and Karl B. Manrodt, 2020, *Supply Chain Management.* New York, NY: McGraw Hill.

PROGRESS CHECK

1. What is the difference between IT operations and supply chain?
2. How are the components of operations related?
3. Which branch of operations do you think is the most important for company success?

LO 10.2

Identify representative operations questions addressable by business analytics in each branch of operations.

10.2 SPECIFYING THE OPERATIONS QUESTION

This chapter examines the three different branches of analytics associated with operations: HR analytics, IT operations analytics, and supply chain analytics.

HR analytics (also known as *people analytics, workforce analytics,* and *talent analytics*) analyzes human resource data to improve a company's workforce performance. HR analytics allows companies to attract, manage, and retain employees to increase company efficiency and profitability, helping managers hire the right employee with the right skill set to the right job at the right compensation level.

IT operations analytics analyzes company data from various applications, devices, and IT infrastructure to help managers identify, predict, solve, and prevent IT-related issues. IT operations analytics helps companies deliver the right IT services at the right quality level with the right level of security at the right time.

Supply chain analytics analyzes company data to help managers improve supply chain planning, efficiency, and effectiveness. Supply chain analytics can help companies integrate, optimize, and maintain supply chain relationships, helping companies deliver the right product at the right time at the right place.

All three types of analytics are designed to create data-based, actionable insights to help companies optimize their decision outcomes. For example:

- **Sainsbury's**, the second-largest grocery chain in the United Kingdom, applies innovations in human resources to continuously generate profits, which is remarkable for an industry characterized by thin profit margins. Sainsbury's is renowned for its employee training academies, and its human resources department pioneered the "self-service style" of shopping that has grown in popularity across the globe.[6]
- **Domino's Pizza**, the world's largest pizza-delivery chain, uses IT operations to help ensure that the company's systems work properly and securely. Half of all Domino's orders are now placed digitally. IT operations is even more critical on its busiest day of the year, Super Bowl Sunday, when 2 million pizzas are ordered (most during a 45-minute window).[7]
- **Walmart**, the world's largest company based on revenue, has developed one of the most technologically advanced supply chain systems. It allows the company to know at all times what products are needed, how many products are needed, and when those products are needed.[8]

[6]Source: HumanResourcesMBA Staff, "The World's 30 Most Innovative Corporate Human Resources Departments." Available online at https://www.humanresourcesmba.net/worlds-30-innovative-corporate-human-resources-departments/. Accessed 3/03/2022.
[7]Source: Splunk, "Revealing the Secret Sauce." Available online at https://www.splunk.com/en_us/customers/success-stories/revealing-the-secret-sauce.html. Accessed 10/08/2021.
[8]Source: Chad Rubin, January 4, 2020, "Walmart Supply Chain 2021: Why It Continues to Dominate?" Available online at https://www.skubana.com/blog/walmart-leading-way. Accessed 10/08/2021.

All three branches of operations analytics begin by specifying a question. Given their different functions and responsibilities, each branch has different types of questions that are potentially answerable by data. Exhibit 10.2 provides examples of detailed operations questions both by analytics type (descriptive, diagnostic, predictive, and prescriptive) as well as by the three main branches of operations: human resources, IT operations, and supply chain.

	Operations Branch		
Business Analytics Type	**Human Resources**	**IT Operations**	**Supply Chain**
Descriptive What happened? What is happening?	How many employees does the company have (head count)? What is the head count in each department? What is the employee turnover rate? On average, how long do employees work for the company? How many days of training does each employee attend in a year?	What is the average transaction response time? What is the average end-user response time for a specific application? What is the application error rate?	Which vendors are the top suppliers (in terms of dollars and units)? How many B2B (business-to-business) transactions did the company have last month? How much inventory is currently on hand? What is the average supplier lead time (to deliver)?
Diagnostic Why did it happen? What are the causes of past results?	Why has head count changed from the past year? Why has employee turnover rate increased? Why has the percentage of candidates hired from the college campus recruiting pipeline fallen?	Why did average transaction response time increase? Why did the application error rate go up/down? How/Why has application performance changed?	Why did quantity used per product change? Why has inventory on hand changed? Why did product quality increase/decrease?
Predictive What will happen in the future? What is the probability something will happen? Can we forecast what will happen?	Which employees have the highest probability of quitting? Which new hires will perform the best? Which teams will underperform due to a skills gap?	Which system will have performance issues? Which system will have security issues? If a data breach occurs, what will happen to response time or cart checkouts?	What will product demand be in two years? What happens to costs if the company increases/decreases order quantities? Which machine will break down?
Prescriptive What should we do, based on what we expect will happen? How do we optimize our performance based on potential constraints?	How many employees does the company need during the holiday season? What is the best way to on-board and train a new hire based on that person's preexisting experience and skills? Where should employees be positioned in a store to improve sales?	Which solution will solve a current systems problem? Which solution should be automatically executed to prevent a systems problem? Based on our company's risk tolerance, how should we set up our systems?	Where should manufacturing plants and warehouses be located? What is the best route to deliver products? Which vendor should we use to meet lead time and quality KPIs (key performance indicators)? Which solution will solve a current machine breakdown?

Exhibit 10.2
Example of Operations Questions by Business Analytics Type and Branch of Operations (Human Resources, IT Operations, Supply Chain)

OPERATIONS ANALYTICS AND YOUR CAREER

One job in operations analytics is the logistic analyst. Logistic analysts monitor the flow of inventory, vehicles, and containers, tracking inventory flow from origin to final destination. In addition, they monitor and communicate with any third parties that ship the company's inventory or parts, including ocean carriers, trucking companies, and air freight forwarders. The median salary is $76,270.[9] Another potential job is HR analytics specialist, whose duties include determining the amount of revenue generated by each employee, measuring voluntary and involuntary turnover, monitoring training costs and effectiveness, and helping to improve the company's hiring strategies. The average salary is $99,680, with top earners making more than $200,000 per year.[10] In IT operations, business analysts predict and prepare for increased Web traffic, identify unusual activity that may signal a security breach, and look for unusual employee (keystroke) behavior that may indicate fraud. The average salary is $67,184, with top earners making more than $100,000 per year.[11] Clearly, there are many great opportunities for people with skills in operations analytics!

PROGRESS CHECK

4. Give two examples of human resource questions that can be answered with HR analytics.
5. Give an example of a supply chain question associated with predictive analytics.
6. As noted in this section, over 2 million pizzas are ordered from **Domino's Pizza** during the Super Bowl. Describe what types of questions Domino's Pizza should ask to prepare for Super Bowl Sunday.

LO 10.3

Identify internal and external sources of operations data.

10.3 OBTAIN THE DATA: OPERATIONS DATA SOURCES

"To be number one, you need to generate innovation. The only way to drive that innovation is based on data."

— ALEX PADILLA, VICE PRESIDENT OF INFORMATION SECURITY, **Domino's Pizza**[12]

A huge amount of data is available to help answer the operations questions in Exhibit 10.2 and many other operations-based questions. In this section, we identify and explain the key sources of data for each branch of operations.

Human Resources Data

Human resources data are potentially created even before a person applies for a position with a company. When a person has an account on **LinkedIn** or posts a résumé on a jobsite

[9]Source: O-Net Online, 2021, *Summary Report for 13-1081.02—Logistics Analysts.* Available at https://www.onetonline.org/link/summary/13-1081.02. Accessed 10/08/2021.
[10]Source: Human Resources MBA, "10 Best Careers for Human Resources Specialists." https://www.humanresourcesmba.net/careers/best-careers-for-human-resources-professionals/#HR%20Analytics%20Specialist. Accessed 10/08/2021.
[11]Source: ZipRecruiter, 2022, "How Much Do IT Operations Analyst Jobs Pay per Year?" Available at https://www.ziprecruiter.com/Jobs/IT-Operations-Analyst?tsid=101050082. Accessed 5/11/2022.
[12]Source: Splunk, "Revealing the Secret Sauce." Available online at https://www.splunk.com/en_us/customers/success-stories/revealing-the-secret-sauce.html. Accessed 10/14/2021.

(such as **indeed.com** or **monster.com**), companies almost immediately have access to these data to help them select qualified applicants for open positions. More data are available if a person officially applies for a position, sits for an interview, or has accessible (and therefore harvestable) social media data. Potential data elements include:

- Education
- Demographics
- Work history
- Certification
- Personality and behavioral traits
- Special skills
- Likes and dislikes

Once a person is hired, the company can collect even more data about the employee, including:

- Work attendance
- Completed training
- Performance evaluations
- Work/job assignment
- Salary
- Promotion history
- Length of employment

To improve the employee experience and predict/reduce turnover, the HR department may also survey employees to get feedback about employee satisfaction, morale, and engagement (that is, how valued employees feel by the company). In addition, data may be collected about the company's communications, employees' experience with the company (which may identify areas for improvement), company culture, employee motivation, and reasons for leaving the company.[13]

Companies also collect and analyze an employee's data trail. As employees perform their jobs they use computer systems, operate office/factory machinery, open and close doors, walk around the company, and interact with other employees and customers. Employees' interactions with the computer system can be captured using keylogging software, activity logs, and automatic screenshots. Companies may also use emails to and from employees to determine how employees across the company interact. In addition, companies may use **telemetry** to collect data at remote points (possibly by a digital employee badge) and automatically transmit the data to monitoring equipment. Telemetry allows companies to understand where employees spend their time, where they go, what they do, and how long they interact with other employees and customers. For example, **Amazon** has a patent for wearable wristbands that can track an employee's location in a warehouse as well as an employee's hand movements. The wristbands can send a vibrating pulse if an employee is not using the correct inventory bin location.[14] Exhibit 10.3 shows an example of a telemetry system.



Although telemetry may provide valuable information to help make the company more efficient, it also raises privacy and ethical concerns about just how much and what types of data should be collected about employees. Should employees have the right to decide

[13]Raymond A. Noe, John R. Hollenbeck, Barry Gerhart, and Patrick M. Wright, 2022, *Fundamentals of Human Resource Management* (9th Ed.). New York, NY: McGraw Hill.

[14]Source: Ally Marotti, 2018, "Amazon Wristbands Could Track Workers' Hand Movements: 'Employers Are Increasingly Treating Their Employees Like Robots.'" *Chicago Tribune* (February 20). Available at https://www.chicagotribune.com/business/ct-biz-amazon-wristbands-privacy-20180215-story.html. Accessed 3/10/2022.

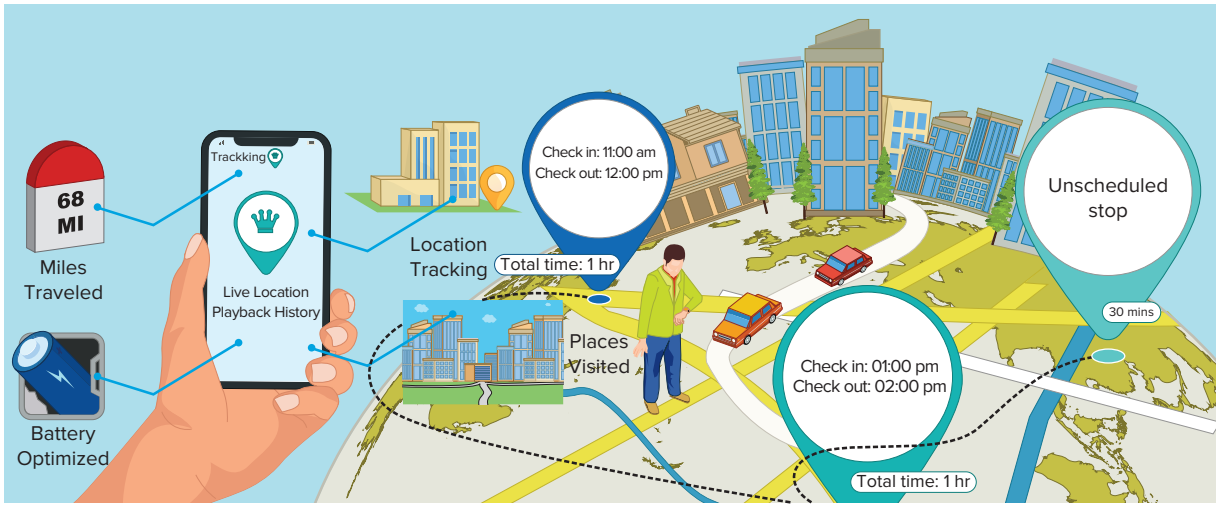


Exhibit 10.3
Sample Telemetry System

Source: Adapted from https://trackking.in/wp-content/uploads/2016/06/employee-tracking-system-1030x343.jpg. Accessed 10/14/2021.

which types of data are collected about them? Or should employers be allowed to collect any data they want? What do you think?

To help manage human resource data, companies may use a **human resource management (HRM) system.** HRM systems typically store all of the collected human resource data in a single place. They help automate the company's human resource processes and compliance with laws and regulations, and they are typically part of the company's enterprise resource planning (ERP) system. (For a review of ERPs, see Chapter 3.) Having a centralized repository of HR data improves data quantity and quality and streamlines the HR analytics process.

IT Operations Data

IT operations collects many types of data from a wide variety of tools and devices to monitor both system performance and service performance, with the goal of improving both. S*ystem performance data* focus on computer applications, devices, and operating systems. For example, a company may monitor and analyze employees' online and company system behavior via usage logs. If employee behavior changes (without a corresponding role change), then the employee may be engaging in negligent or illegitimate actions that should be further investigated. Likewise, companies monitor and analyze network traffic for changes that may point to a security breach. *Cybersecurity analytics* analyzes historic system data to pinpoint the breach source and to identify the systems, programs, or accounts that were compromised. Potential data elements used for system performance analysis include the following:

- User behavior via usage logs
- Performance metrics, such as level of usage and response time
- Data breaches
- Network information, such as data about information flows between applications and servers to identify source and destination machines/devices

Service performance data focus on the help desk, change management, and incident response to IT problems. For example, a company may want to make sure that all IT help desk employees are quickly responding and resolving employees' IT problems in a professional manner. Therefore, the company may track how long it takes the help desk to

respond to a service request, how long it takes to resolve the problem, and the employee's satisfaction with the interaction and final resolution. The company will then analyze the data by help desk employee to make sure they are all working effectively and efficiently. If certain employees are underperforming, the company could offer additional training to them. Potential data elements used for service performance analysis include the following:

- Incident-related data
- Service/support ticketing—that is, all service desk interactions with customers and employees
- Service metrics, such as response time[15]

If the company outsources part or all of its IT to a vendor, then the company most likely has a **service-level agreement (SLA)** with the vendor. An SLA defines the level of service that the company expects from the technology vendor/provider as well as penalties if the agreed-on service levels are not attained. For example, a technology vendor/provider, such as **Oracle** for expansive enterprise systems or **Patriot Software** for accounting and payroll software, may promise 99.999 percent availability. If this service level is not attained, then the company gets a 10 percent reduction in the monthly rate paid to the technology vendor for each instance of noncompliance. The items tracked to evaluate the SLA depend on the services being provided but may include the following:[16]

- Service availability
- Defect rates, such as the number of incomplete backups
- Technical quality of outsourced development
- Security issues, including breaches, updates, patches, antivirus protection, and firewalls
- Vendor's contribution to specified company performance measures

Supply Chain Data

Recall that the supply chain incorporates all of the processes required to procure the raw materials, manufacture the product, and deliver the product to its final destination. Therefore, supply chain analytics tends to require more data sources than other types of analytics. To help collect data, a company can use a supply chain management **(SCM) system**, which is a computer system that collects and stores supply chain data in one place. It helps the company automate supply chain processes and comply with laws and regulations. SCM systems can also help forecast demand, order raw materials to avoid stockouts (the complete depletion of inventory), track the raw materials from shipping through the manufacturing process, monitor warehouse storage, and track shipping to the final customers. Some companies also use warehouse management systems to manage warehouse operations. Moreover, a company can link its SCM system with its customer relationship management (CRM) system with the goal of enhancing customer experience and satisfaction.[17]

[15]For more information, please see https://www.ibm.com/cloud/learn/aiops and https://www.gartner.com/en/documents/3892967/market-guide-for-aiops-platforms. Accessed 10/14/2021.

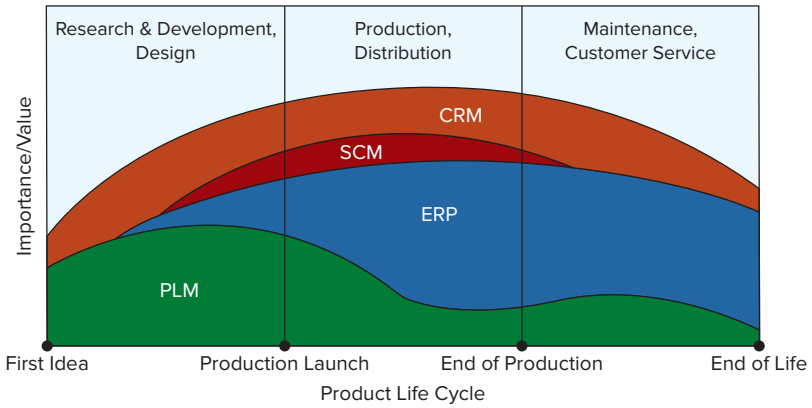
[16]NetSuite and Patriot Software are used only as examples. The exact details of SLAs with customers are not known. The list of potential data items collected comes from Stephanie Overby, Lynn Grenier, and Lauren Gibbons Paul, "What Is an SLA? Best Practices for Service-Level Agreements," *CIO.com,* July 5, 2017, https://www.cio.com/article/2438284/outsourcing-sla-definitions-and-solutions.html. Accessed 10/14/2021.

[17]A CRM (customer relationship management) system manages the company's interactions with former, existing, and potential customers. It also stores contact information, records sales and service issues, helps the company manage marketing campaigns to identify potential sales opportunities, and manages/records communications with customers across marketing channels. See Chapters 4 and 7 for more information.

As Exhibit 10.4 shows, CRM, ERP, and SCM all play an important role over a product's life cycle. In the exhibit, PLM means product life cycle management, which refers to how the company manages the product over the product life cycle. The *product life cycle* is a series of steps or stages through which all products go. Companies need to focus on different items in each stage. The relative size of the color band indicates the value of different systems over the product life cycle. So, as Exhibit 10.4 shows, PLM and ERP play the largest roles throughout a product's life cycle. SCM plays an increasing role at the end of the Research & Development, Design phase; a larger role during the Production, Distribution phase; and a diminishing role in the Maintenance, Customer Service phase. CRM plays a consistent role throughout the product life cycle, but its role is not as large as PLM and ERP.

Exhibit 10.4
Flow Graph of CRM, ERP, and SCM

Source: Adapted from EccelentSolutions. "How do you 'PLM'?" Available online at https://www.eccel-lent.com/how_do_you_plm/. Accessed 3/2/2022.



CRM: Customer Relationship Management system
ERP: Enterprise Resource Planning system
SCM: Supply Chain Management system
PLM: Product Life Cycle Management

Suppose a company is designing a new product. In the Research & Development, Design phase it may use computer-aided design coupled with feedback on current products from the CRM system. As the company builds prototypes of the product, it needs parts from vendors. These vendors will be set up in the SCM system to allow the company to procure the parts moving forward. The ERP system will initially need part numbers and cost figures set up for the new product to pay the vendors and to record sales after the product is launched. Assuming the product is a success, a large volume of sales transactions will be recorded in the ERP system, and customer information will be tracked in the CRM. When the product is discontinued, the focus will be on servicing products, with the costs being recorded in the ERP system. PLM integrates information from all of the other systems throughout the product's life cycle, increasing efficiency and reducing costs.

Companies also harness data throughout the supply chain (warehouses, delivery trucks, machines, and store shelves) using a variety of different sensors (telemetry). For example, companies can automatically count inventory in the warehouse and perform real-time monitoring of delivery trucks to determine location, stops, speed, and even the temperature of the cargo. Technology can also be used to monitor company assets, such as trucks and machines, for maintenance needs.

Companies may want to augment their internal supply chain data with external data. For example, weather and traffic data can help determine delivery times, route delivery trucks, and predict customer demand. Consumer feedback on social media provides valuable inputs to forecast demand and develop product features and design. Data about customers' and suppliers' financial health are critical for demand planning as well as continuing raw-materials acquisition. Finally, data from the government help ensure that the

company's supply chain is always in compliance with the ever-changing governmental rules and regulations.

Here is a list of other data that may be collected for supply chain analytics:

- Manufacturing, shipping, and warehouse capacity
- Inventory levels, including stockouts (complete depletion of inventory)
- Radio-frequency identification (RFID) tracking data
- Shipping metrics, such as on-time arrivals, delays, and accuracy
- Demand forecasts
- Supplier performance metrics, such as on-time delivery, price, quality, and service
- GPS information
- Maintenance logs and alerts
- Raw material prices

PROGRESS CHECK

7. Identify sources of internal data that can be used for operations analytics.
8. Identify sources of external data that can be used for operations analytics.
9. Identify sources of data that **Domino's Pizza** may use to prepare for the upcoming sales on Super Bowl Sunday.

10.4 DESCRIPTIVE OPERATIONS ANALYTICS

LO 10.4

Describe and provide examples of descriptive analytics used to answer operations questions.

After the operations questions and appropriate data have been identified, companies are ready to perform operations analytics. In this chapter, we focus on four types of operations analytics: descriptive, diagnostic, predictive, and prescriptive. We begin with descriptive analytics.

Recall that descriptive analytics describes or summarizes the past. It helps highlight the characteristics and features of a data set by summarizing a large amount of data in an easy-to-understand format. The results of descriptive analytics do not include any assumptions or future estimates/predictions. In other words, descriptive analytics helps us understand the *actual* data and what has already happened.

Descriptive analytics uses a wide variety of summarization and statistical techniques as well as simple visualizations. We discuss some examples in the following sections.

Statistical, Summarization, and Data Visualization Techniques

Data analysts use summarization and statistical techniques to answer the questions "What happened?" and "What is happening?" These techniques characterize, summarize, and organize data and are generally available in spreadsheet software such as Excel. For example, counts, totals, sums, averages, subtotals, minimums, maximums, pivot tables, standard deviations, and ratios are all descriptive analytics that Excel can generate. In addition, business analysts can use Excel to create simple charts and visualizations such as histograms, bar charts, and pie charts. Exhibit 10.5 provides examples of statistical techniques that can be used to answer questions about the different branches of operations. The term **fill rate** in the exhibit means the percent of customer orders satisfied by inventory on hand.

		Example Operations Question		
Statistical Technique for Descriptive Analytics	**Purpose**	**Human Resources**	**IT Operations**	**Supply Chain**
Counts	Determine frequency of occurrence	How many new employees were hired during the year?	How many times did the website crash during the last month?	How many times did the company stockout (that is, run out of inventory)?
Graphs: bar charts, histograms	Provide summary visualization	How many employees are assigned to each location?	What is the average number of users on each system for a month?	What is the average lead time for each supplier?
Minimums, maximums, medians, standard deviations	Provide summary measures of dispersion	Which employee has the lowest salary?	Which employee spends the most time on the company's enterprise system?	Which supplier has the highest frequency of missing scheduled delivery times?
Pivot tables	Provide a flexible way to summarize large amounts of data	What is the total number of employees by month and location?	What is the total number of users by each system and by month?	What is the number of error-free orders by day and shift?
Ratios	Compare two numbers	What is the employee turnover rate for the year?	What is the ratio of open versus completed service tickets?	What is the supply chain fill rate?
Totals, sums, averages, subtotals	Provide summary measures of performance	How many total employees work for the company?	What is the average website response time?	What is the average supplier lead time?

Exhibit 10.5
Statistical Techniques for Operations Descriptive Analytics

Head Count

"Your workforce is your most valuable asset. The knowledge and skills they have represent the fuel that drives the engine of business—and you can leverage that knowledge."

— HARVEY MACKAY, BUSINESSMAN, **New York Times** best-selling author, and columnist

Descriptive analytics provides the basic statistics for operations data. Many governmental entities provide data about their operations on a government website. Using the *open data,* we can evaluate the operations of government entities.

For example, using data from the U.S. government, we can answer this question: Which (U.S.) federal agency employs the most people? Exhibit 10.6 shows the top six federal agencies based on total number of employees. It reveals that, by far, most of the federal employees work for the Veterans Health Administration (VHA), followed by the Internal Revenue Service (IRS), Air Force Materiel Command, Transportation Security Administration (TSA), Social Security Administration, and Customs and Border Protection.[18] Public and private companies can perform a similar analysis of head counts by department, region, product, or some other relevant dimension.

Website Response Time

Descriptive analytics helps analyze website response and downtimes; IT help-desk response and resolution times; and the activity of employees, customers, and hackers.

[18]The Department of Defense has its own data available on the same website, but these data are not included in Exhibit 10.7.

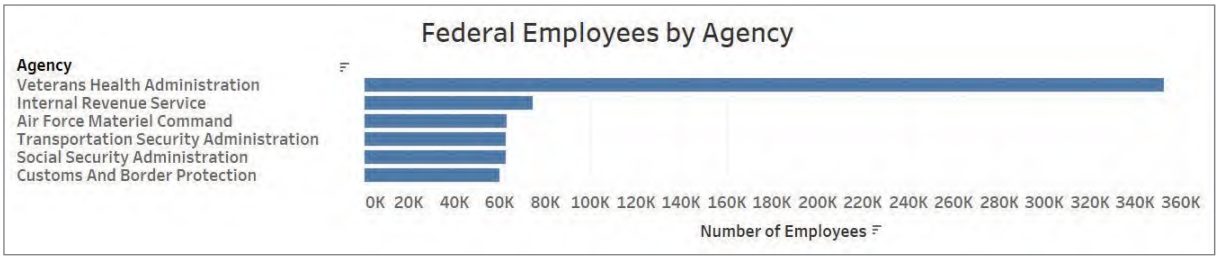


Exhibit 10.6
Number of Employees by Federal Government Agency

Source: https://www.fedscope.opm.gov. Accessed 7/18/2020.

MINI CASE: THE CITY OF CHICAGO's GOAL TO BE ONLINE

The City of Chicago maintains six websites (for example, https://www.chicago.gov/city/en.html) to provide resources and information about the city. The City automatically tracks the response times for the six websites to help ensure that all websites are working properly. We can graph the website response times using a **time series graph**, which is a visual representation of a variable over time. A times series graph shows how a variable changes over time. Exhibit 10.7 shows a time series graph of almost 55 weeks of data from the Chicago Data Portal.

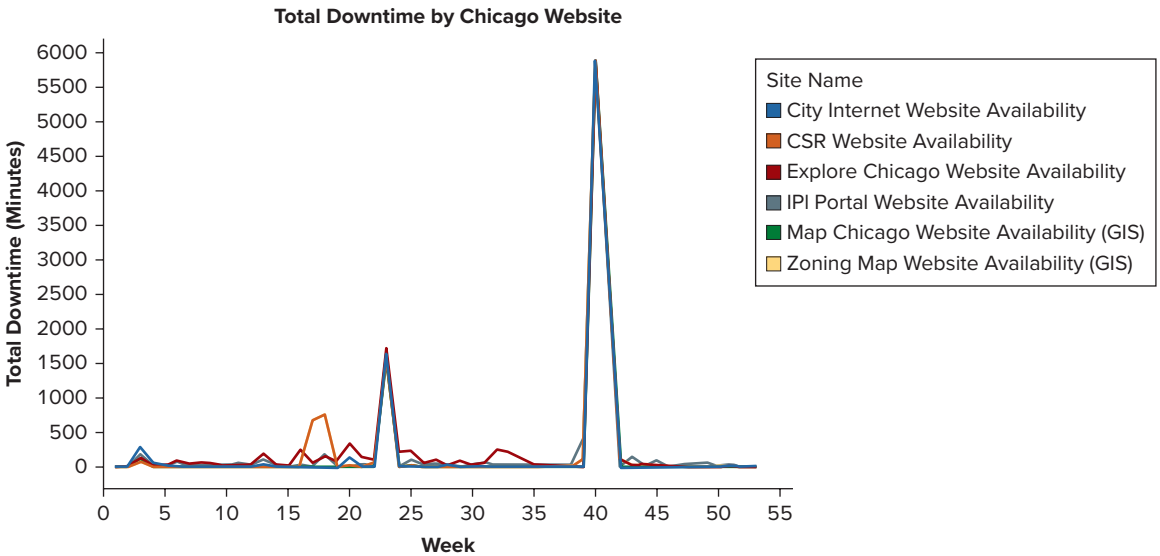


Exhibit 10.7
Total Downtime for Chicago Websites

Data Source: https://data.cityofchicago.org/Administration-Finance/Performance-Metrics-Innovation-Technology-Site-Ava/zfg3-p7xv. Accessed 11/03/2021.

Decision-Making with Data

How can the City of Chicago use Exhibit 10.7 to answer the following questions: How many minutes of downtime do the websites experience each week? Does downtime vary over time? Do certain websites have more downtime than other websites? Based on the analysis, what additional analytics should the City of Chicago perform?

Key Performance Indicators (KPIs)

An integral part of both descriptive and diagnostic analytics is **key performance indicators (KPIs),** which are numerical metrics used to gauge company performance. Each company determines which KPIs are the most crucial for the business and how they should be broken down—for example, by channel, product, region, system, or store. The following KPIs are commonly used in operations.

- Human resources:
  - Absence rate and cost
  - Employee productivity
  - Employee job and benefits satisfaction
  - Internal promotion rate
  - Involuntary and voluntary turnover rate (percent of employees no longer employed)
  - Training effectiveness and cost per employee
  - Total employee cost as a percentage of total costs
  - New hire cost and 90-day failure rate
- IT operations:
  - Open help-desk tickets versus resolved tickets
  - Server and website downtime
  - Unsolved tickets per employee
  - Reopened tickets
  - IT support employees versus total employees
  - Server and website response time
  - Average incident resolution time by type of incident (for example, application, cyber)
  - Server and website utilization rates
  - SLA agreed-upon measures
  - Intrusion attempts
  - Security incidents (breaches)
  - Average time to detect security threat
- Supply chain:
  - Cash-to-cash cycle time (*conversion cycle*) and its components: days of payables, inventory, and receivables
  - On-time delivery to customer
  - Inventory-to-sales ratio
  - Shipping cost and time
  - Perfect order delivery rate
  - Slow-moving stock
  - Supplier on-time delivery and quality
  - Inventory turnover and carrying cost
  - Rate of product returns by customer
  - Back-order rate

Let's look at an example using KPIs. Specifically, let's analyze a variety of supply chain ratios for retail and wholesale trade companies over the past decade. Keep in mind that these companies are not manufacturing companies. Rather, retail and wholesale trade companies buy finished goods from a manufacturer, store the goods in warehouses or some other location, and then sell the goods to customers. We want to answer the following questions:

(1) How efficient is the supply chain?
(2) How responsive is the supply chain?
(3) How reliable is the supply chain?

To answer the first question about supply-chain efficiency, we need to calculate the following two ratios:

(a) the *cash-to-cash cycle,* which measures the length of time between the company paying its vendors and the company receiving cash from its customers. The lower the ratio, the better.

Cash-to-cash cycle = Days of inventory + Days of accounts receivable – Days of accounts payable

where

- Days of inventory = number of days a company holds the item in inventory before it is sold
- Days of accounts receivable = number of days that it takes a customer to pay for the item
- Days of accounts payable = number of days it takes for the company to pay the vendor for the item

(b) *asset turnover,* which measures the sales revenue generated by the company's assets. The higher the ratio, the higher the efficiency.

$$\text{Asset turnover} = \frac{\text{Sales revenue}}{\text{Total assets}}$$

To answer the second question about supply chain responsiveness, we need to calculate the *inventory-to-sales ratio,* which measures the amount of inventory being carried to fulfill sales. A high inventory-to-sales ratio suggests that the company is spending more money on storage and handling costs. So, low is the way to go!

$$\text{Inventory-to-sales} = \frac{\text{Inventory}}{\text{Sales}}$$

To answer the third question about supply-chain reliability, we need to calculate the *inventory turnover ratio,* which measures the number of times that inventory is sold during the year. A higher inventory turnover ratio is typically desired because it indicates reduced storage and handling costs.

$$\text{Inventory turnover} = \frac{\text{Cost of goods sold}}{\text{Inventory}}$$

Exhibit 10.8 presents the ratios generated using the Descriptive Statistics menu option of Excel's Analysis ToolPak. On average, the cash-to-cash cycle is 56 days (essentially two months), each dollar of assets generates two dollars of sales revenue, very little inventory is being held to cover sales, and inventory is completely sold approximately 23 times a year (meaning that on average it takes about 16 days, or 365/23, to sell inventory). The descriptive statistics also reveal that some of the ratios have great variability. A more in-depth analysis, perhaps by industry, could reveal even more information to evaluate the efficiency of different industries' supply chains.

LAB CONNECTIONS

LABS 10.1 and **10.2** provide step-by-step instructions on how to generate descriptive analytics using Excel, Tableau, and Power BI.

- **LAB 10.1** shows how to use Excel to generate descriptive analytics for supply chain KPIs.
- **LAB 10.2** shows how to use Tableau and Power BI to create bar charts for human resource data.

Exhibit 10.8
Sample Supply Chain KPIs for Retail and Wholesale Trade Industries, 1999-2021

Source: Wharton Research Data Services (WRDS). Compustat North American Fundamentals-Annual. Accessed 3/30/2022.

	Efficiency		Responsiveness	Reliability
	Cash-to-Cash Cycle	*Asset Turnover*	*Inventory-to-Sales*	*Inventory Turnover*
Mean	55.71	2.05	0.13	23.20
Standard Error	1.06	0.02	0.00	0.68
Median	40.41	1.77	0.11	6.48
Mode	15.70	0.06	0.08	227.58
Standard Deviation	86.03	1.71	0.15	55.25
Sample Variance	7,401.89	2.91	0.02	3,052.23
Kurtosis	28.95	325.38	211.86	135.33
Skewness	3.31	11.36	9.32	9.05
Range	1,549.73	66.00	4.93	1,271.38
Minimum	(506.15)	0.00	0.00	0.34
Maximum	1,043.58	66.00	4.93	1,271.71
Sum	368,647.69	13,579.62	862.7	153,542.18
Count	6,617	6,617	6,617	6,617

PROGRESS CHECK

10. Define the following terms used in the supply chain descriptive statistics example above: *days of payables, days of inventory,* and *days of receivables*.
11. What operations KPIs would be the most important for **Domino's Pizza** on Super Bowl Sunday?

LO 10.5

Describe and provide examples of diagnostic analytics used to answer operations questions.

10.5 DIAGNOSTIC OPERATIONS ANALYTICS

Diagnostic analytics helps answer the questions "Why did it happen?" and "What are the causes of past results?" In other words, diagnostic analytics attempts to identify the root cause —the factor that initiated the chain of events—which usually cannot be uncovered by descriptive analytics. Diagnostic analytics helps evaluate the relations, patterns, and linkages between variables that might be related to an anomaly. Performing diagnostic analytics often requires combining multiple data sets and sometimes using both external and internal data.

Identifying Anomalies/Outliers

Often a first step in diagnostic analytics is to look for and identify unusual, unexpected results, known as anomalies or outliers. Exhibit 10.9 details some anomalies by branch of operations and provides some possible explanations.

Branch of Operations	Expectation	Example of Anomaly	Possible Explanations for the Anomaly
Human Resources	Annual employee turnover will be 5%.	Annual employee turnover is 15%.	Company salaries are below the industry average.
IT Operations	Monthly server downtime will be 30 minutes.	The server downtime for the month of May was 2 hours.	A patch to fix server issues was not applied.
Supply Chain	Average freight cost per ton-mile will be $18.80.	Average freight cost per ton-mile for September was $30.66.	More freight was shipped using airplanes instead of trucks.

Exhibit 10.9
Examples of Anomalies by Operations Branch

Once the unexpected results have been identified, a company can investigate to understand exactly why those results occurred. As described in Chapter 7, root cause analysis, potentially with an Ishikawa Diagram, is often initially used to help understand what is driving the unexpected change. (See Learning Objective 7.5.) Additional techniques like those identified in Exhibit 10.10 can be used to find patterns or linkages in the internal or external data set.

		Example Operations Questions		
Statistical Technique for Diagnostic Analytics	**Use**	**Human Resources**	**IT Operations**	**Supply Chain**
Control Chart	Determines whether a process is in control.	When and how frequently is employee productivity not in compliance with company standards?	When and how frequently is system response time not in compliance with company standards?	When and how frequently is vendor quality not in compliance with company standards?
Correlation	Determines the extent to which variables are related to each other.	To what extent is employee satisfaction related to customer satisfaction and sales revenue?	To what extent is system response time related to the time of day?	To what extent is on-time delivery rate related to the season?
Hypothesis Testing	Tests whether variables are significantly different from each other.	Has the percentage of employees voluntarily leaving the company changed?	Has the number of minutes of downtime for websites increased?	Have stockouts decreased?
Pivot Table/ Cross-Tabulation	Groups variables to understand the relationships between them. Allows drill down and drill up.	What is the average length of employment for different types of employees?	What is the average resolution time for a systems problem in total and by each IT staff member?	What is the average freight cost by delivery method?
Regression	Examines whether there is a relationship between an independent input variable and dependent (outcome) variable.	How is training related to employee turnover?	How is IT service-ticket resolution related to employee satisfaction with IT operations?	How is weather related to demand?
Summarization	Shows how totals vary at different levels of granularity.	Why has employee turnover increased? Is it a company-wide issue or localized issue?	Why has average system response time increased? Is it a company-wide issue or a specific system issue?	Why has average delivery time increased? Is it widespread or just in certain cities?

Exhibit 10.10
Diagnostic Analytics Techniques and Examples for Operations

Hypothesis Testing

Exhibit 10.8 presented some average supply chain KPIs during the 1999–2021 period. Over the past decades, technology has permeated and transformed the supply chain, making it reasonable to assume that these ratios may have changed, and improved, over time as technology increased the efficiency of the supply chain processes. To investigate if this hypothesis is true, we want to answer the question: Have the supply chain KPIs changed over time?

To answer this question, we will perform hypothesis tests comparing the ratios from 1999 to those in 2021 using the t-Test: Two Sample Assuming Unequal Variances menu option of

the Data Analysis ToolPak for Excel. (See Chapter 3 for an in-depth discussion of hypothesis testing.) A two-sample option is used because we compare all of the companies in 1999 to those in 2021, and the number of companies available each year is different. Exhibit 10.11 presents the results of the hypothesis tests comparing the KPIs for cash-to-cash cycle, asset turnover, inventory-to-sales ratio, and inventory turnover for 1999 and 2021.

Let's look at the two-tailed test because we did not predict a specific direction with our question. Small values (with most people using a 0.05 cutoff) indicate that the variable changed significantly. Exhibit 10.11 shows that cash-to-cash cycle, asset turnover, inventory-to-sales ratio, and inventory turnover changed significantly. Specifically, cash-to-cash cycle significantly decreased from 69 to 44 days. The asset turnover ratio significantly decreased from 2.13 to 1.60, indicating that assets are generating fewer sales today than they did in 1999. The inventory-to-sales ratio significantly decreased from 0.15 to 0.12, indicating that the amount of inventory held (and associated costs) to fulfill sales decreased. Inventory turnover increased from 15.74 to 29.69, indicating that inventory was being sold faster in 2021 than in 1999 (potentially from supply chain improvements). Additional hypotheses tests could be performed on the components of the ratios to pinpoint exactly what changed.

t-Test: Two-Sample Assuming Unequal Variances

	Cash-to-Cash Cycle		Asset Turnover		Inventory-to-Sales		Inventory Turnover	
	1999	*2021*	*1999*	*2021*	*1999*	*2021*	*1999*	*2021*
Mean	68.91	43.76	2.13	1.60	0.15	0.12	15.74	29.69
Variance	10028.14	3620.24	1.93	0.92	0.02	0.01	826.14	6895.48
Observations	207.00	258.00	207.00	258.00	207.00	258.00	207.00	258.00
Hypothesized Mean Difference	0.00		0.00		0.00		0.00	
df	321.00		352.00		346.00		330.00	
t Stat	3.18		4.64		3.19		-2.52	
P(T<=t) one-tail	0.00		0.00		0.00		0.01	
t Critical one-tail	1.65		1.65		1.65		1.65	
P(T<=t) two-tail	0.00		0.00		0.00		0.01	
t Critical two-tail	1.97		1.97		1.97		1.97	

Exhibit 10.11
Sample Supply Chain KPIs for Retail and Wholesale Trade Industries, 1999–2021

Yellow highlights significant differences at the 0.05 level.
Source: https://wrds-www.wharton.upenn.edu/ Research Data Services (WRDS). Compustat North American Fundamentals-Annual. Accessed 3/30/2022.

Control Charts

Control charts are graphs that plot data about a process in a time-ordered sequence. A control chart consists of an average line, an upper control limit line, and a lower control limit line, all determined based on historical data. By comparing current process data to these three lines, companies can determine if a process is in control or out of control.

The upper control limit (UCL) and lower control limit (LCL) are typically set three standard deviations from the mean. Observations are considered outliers (and unacceptable/out of control) if they are outside the designated UCL and LCL. These outliers typically have an underlying cause rather than just randomness. Control charts may also exhibit trends (defined in this example as 6 or more points ascending or descending) and shifts (9 consecutive points above or below the average).

Typically, companies start by trying to minimize data points outside of three standard deviations and then slowly move the acceptable UCL and LCL closer together—for example, moving from 2.5 standard deviations, to 2 standard deviations, to 1.5 standard deviations, and so on. **Six Sigma**, which uses techniques and tools for continual process improvement almost to the point of zero defects, utilizes control charts to identify and reduce variation. Six Sigma is named for the three standard deviations on either side of the mean of a normally distributed variable.

Exhibit 10.12 is an example of a control chart. The solid line plots the observations over time, or the time series. The Average blue dotted line shows the average of all of the observations and serves as the reference point for the UCL and the LCL. In this example, only one point is an outlier for the process: the single point below the LCL, at the right side of the figure. Shifts and trends are also used to identify areas of improvement by Six Sigma.

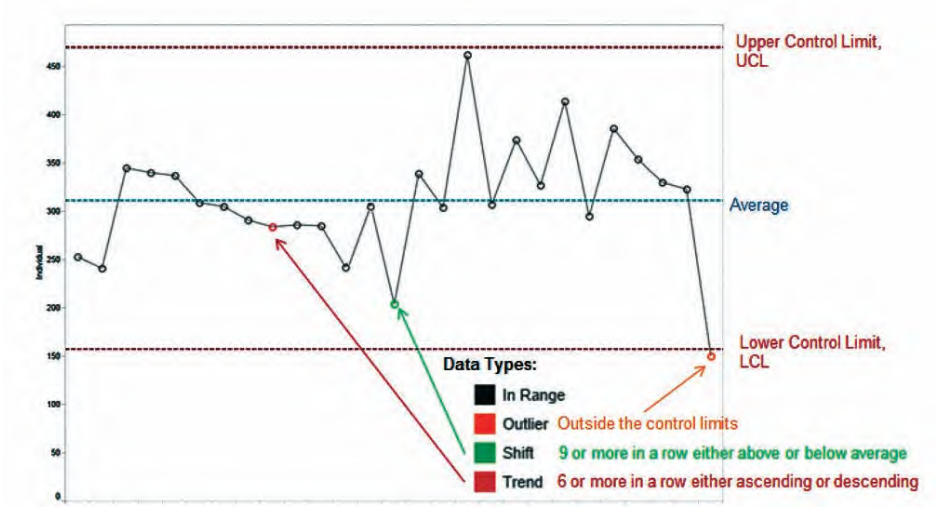


Exhibit 10.12
An Example of a Control Chart

Source: Ben Jones, November 27, 2013, "How to Make Control Charts in Tableau." Available online at https://public.tableau.com/en-us/s/blog/2013/11/how-make-control-charts-tableau. Accessed 10/15/2021.

SOARing to Success

The Origins of the Total Quality Movement

Believe it or not, control charts are 100 years old. Control charts were invented in the 1920s by Walter Shewhart at the **Western Electric Company** to improve the quality of telephone equipment. Shewhart is sometimes called the grandfather of Total Quality Movement because his ideas sparked a movement. W. Edwards Deming championed Shewhart's work and developed 14 points for implementing quality management. Deming is most famous for his work with **Toyota** and **Sony** in Japan after World War II. He helped Japanese companies learn how to manufacture quality goods economically. One of his major ideas was to "bake" quality into the product instead of inspecting for defects after the fact. Eventually, U.S. companies, including the **Ford Motor Company**, **Xerox Corporation**, **Procter & Gamble Company**, **AT&T Inc.**, and the ***New York Times***, hired Deming to help them restore global competitiveness and profitability. Deming is called the Father of the Total Quality Movement and the God of Quality because his ideas transformed companies across the globe.

MINI CASE: FORD MOTOR COMPANY: IS QUALITY REALLY JOB ONE?[19]

In the 1980s, the Ford Motor Company was not only losing market share to Japanese car manufacturers but also losing billions of dollars. To help make the company profitable again, Ford hired W. Edwards Deming. The result of partnership was the 1986 Motor Trend Car of the Year, the aerodynamic Ford Taurus, which helped make Ford profitable

[19]Source: https://towardsdatascience.com/netflix-recommender-system-a-big-data-case-study-19cfa6d56ff5. Accessed 8/18/2021.

again. Ford adopted the slogan "Quality Is Job One" as part of this retooling, baking quality into its products.

Now, let's fast-forward 30 years. Quality apparently is no longer job one at Ford. As Exhibit 10.13 shows, warranty claims for the Ford Motor Company have doubled over the past ten years, rising from two percent to over four percent of sales. One-third of Ford's warranty claims are the results of faulty parts from suppliers. The rest of the warranty claims are the result of production assembly-line problems, glitchy software, and poor/complex design issues. The impact and cost of these problems has been catastrophic. According to **Consumer Reports**, Ford ranks 22nd out of 26 auto brands, with the Lincoln brand at the bottom. For the first nine months of 2020, Ford paid $3.9 billion for warranty claims compared to $1.7 billion at **General Motors**, 3.1 percent versus 1.4 percent of sales revenue, respectively.[20] Ford hopes to quickly reclaim Quality is Job One to improve profitability and consumer perception of its brand.

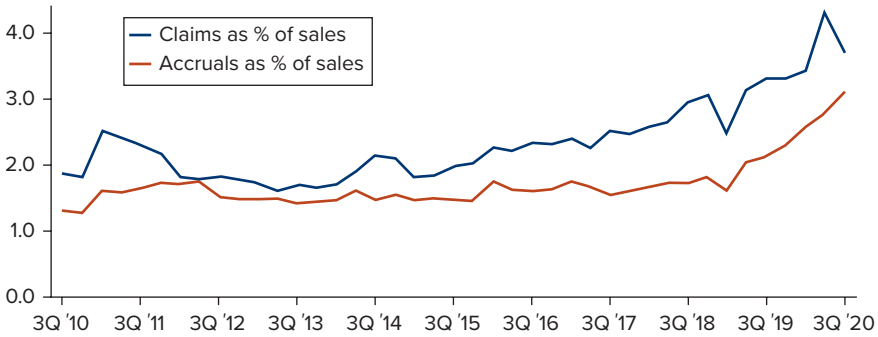


Exhibit 10.13
Ford Motor Company's Warranty Claims
Ford's warranty claims as a percentage of sales doubled between 2010 and 2020.

Source: https://graphics.reuters.com/FORDMOTOR-WARRANTY/qmypmxoyzvr/index.html. Accessed 10/15/2021.

Decision-Making with Data

How could Ford use control charts to help improve product quality? What control charts would you recommend to help minimize supplier, assembly-line, software, and design problems?

LAB CONNECTIONS

LABS 10.3 and **10.4** provide step-by-step instructions on how to conduct two diagnostic techniques.

- **LAB 10.3** shows how to how to perform a hypothesis test in Excel comparing the supply chain KPIs of different companies.
- **LAB 10.4** shows how to create a control chart in Tableau.

[20]Sources: Ben Klayman, November 24, 2020, "Ford's New CEO Tackles Warranty Costs in Bid to Boost Profit." Available online at https://www.reuters.com/article/us-ford-motor-warranty-focus-idUSKBN2841BV. Accessed 10/15/2021. Paul A. Eisenstein, November 24, 2020, "Ford CEO Farley Wants to Make Quality Job 1 Again," The Detroit Bureau, https://www.thedetroitbureau.com/2020/11/ford-ceo-farley-wants-to-make-quality-job-1-again/. Accessed 10/18/2021.

PROGRESS CHECK

12. What is the difference between KPIs used for descriptive analytics and KPIs used for diagnostic analytics?
13. What other diagnostic techniques (other than control charts) could **Ford** use to improve its products?

10.6 PREDICTIVE OPERATIONS ANALYTICS

LO 10.6

Describe and provide examples of predictive analytics used to answer operations questions.

Predictive analytics answers the questions: "What will happen in the future?"; "What is the probability something will happen?"; and "Can we forecast what will happen?" The goal is to predict future performance using historical data and a user-specified **algorithm**, which is basically human-programmed computer instructions for how to combine the data. So, unlike descriptive and diagnostic analytics, which describe things we know are true, predictive analytics describes things that possibly might happen.

There are four broad categories of predictive analytic techniques: (1) classification, (2) decision trees, (3) forecasting, and (4) regression.[21] Exhibit 10.14 outlines these four predictive analytics techniques, summarizes how they are used, and provides an example of an operations question that each technique can help to answer.

		Example Operations Question		
Statistical Technique for Predictive Analytics	**Use**	**Human Resources**	**IT Operations**	**Supply Chain**
Classification	Separates a data set into two or more predefined "like" groups	Which new employees are more likely to perform well?	Which employees are more likely to be engaging in fraud?	Which vendors are more likely to meet company standards in the future?
Decision Tree	Predicts a specific dependent (outcome) variable based on a pathway that includes all action points with associated probabilities	How well will a new hire perform?	Should we develop our own system or use a vendor solution?	How much capacity should our manufacturing plant have?
Forecasting	Predicts future values based on past values of the same variable; anticipates the behavior of many people/things on long timelines	What will employee turnover be next year?	How many active users will be using a specific company system in two years?	What will customer demand for a product be next year?
Regression	Predicts a specific dependent (outcome) variable based on the relationship between independent and dependent variables	Which employees have the highest likelihood of leaving?	Which computer systems have the highest likelihood of breaking down?	Which machines have the highest likelihood of breaking down?

Exhibit 10.14
Predictive Analytics Techniques and Operations Examples

[21]Decision trees can be used to create classification trees (for categorical data) and regression trees (for continuous data).

Regression

Regression analysis uses statistics to determine (1) if variables have an impact on an outcome and (2) the size of the impact. Regression analysis helps answer the questions "Which factors matter the most in predicting an outcome?" and "Which factors can we ignore (because they don't predict the outcome)?" Let's look at a simple regression that answers the question "What is the total manufacturing cost for a given number of units produced?"

Let's think through the relationship first. For example, it seems that the greater the number of units produced, the more likely total costs will increase. Therefore, we expect a positive relationship between units produced and total costs.

To see if there is a statistical relationship between units produced and total cost, we will run a regression with the following specifications:

- Dependent (outcome) variable: Total cost
- Independent variable: Production units

To run the regression, we use the Excel Analysis ToolPak's Regression option. Exhibit 10.15 presents the regression output using monthly data from last year. The model fits the data fairly well because the Adjusted R Square equals 0.77. The regression equation is: Total Cost = 81449.94 + 26.81*Production Units. The intercept represents total fixed cost of $81,449.94, while the slope of the line represents the per-unit variable cost of $26.81. In other words, Total Cost = Fixed Cost + (Variable Cost * Production Units) where:

- Fixed costs are those costs that a company incurs no matter how many units are produced. Examples include insurance, rent, depreciation, and property taxes.
- Variable cost is the cost of producing one unit of product. Examples of variable costs include raw materials cost, direct labor cost, and packaging cost.

We can use this equation to estimate total cost. Assume that we want to know the total cost of producing 1,300 units. Based on the regression equation, we estimate the cost to be $81,449.94 + $(26.81 * 1,300) = $116,302.94. However, this is just an estimate. Even though Adjusted R Square is fairly large, if you look at the lower 95 percent and upper 95 percent confidence intervals for the Intercept and Production Units in Exhibit 10.16, you will see that they are fairly large. For example, we can say with 95 percent confidence that per-unit cost (Production Units) has a value between $17.07 and $36.55. Our estimate above is made using the coefficient value of $26.81. Management will need to decide if the model is accurate enough for its current decision-making process.

Forecasting

As the above example shows, regression predicts a specific dependent-variable outcome based on the relationship between independent and dependent variables. Supply chain management uses regression to forecast product demand to help ensure that the right products are available when needed. Specifically, demand forecasts help prevent the overstocking and understocking of inventory, which can both negatively impact revenue. In addition, demand forecasts can help companies plan for production, purchases, and head count.

Let's assume that a company has four warehouses that fulfill product orders for customers. Each warehouse is responsible for a different region. The company would like to predict how many workers are needed at each warehouse, which is a function of demand. The question we are answering in this analysis is: What is expected demand for 2022 at each warehouse?

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89
R Square	0.79
Adjusted R Square	0.77
Standard Error	3138.86
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	370581582.31	370581582.31	37.61	0.000110777
Residual	10	98524619.69	9852461.97		
Total	11	469106202.00			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	81449.94	6263.73	13.00	0.00	67493.49	95406.40
Production Units	26.81	4.37	6.13	0.00	17.07	36.55

Exhibit 10.15
Operations Regression Example

Our data consist of order date, quantity demanded, and warehouse location for 2017 to 2021.[22] The quantity demanded is a time series because there is an explicit order or time dimension to the observations. In other words, the quantity demanded must be ordered sequentially as it took place. Because this is a time series, we can use the past values of quantity demanded to forecast future values (that is, a time series forecast) and answer our question.

To begin our analysis, we must first understand the data. To do so, we graph the data in Tableau, as shown in Exhibit 10.16. The actual data are represented by solid lines, with each warehouse assigned a different color. Then, using the Forecast menu option (under Analysis), Tableau generates for each warehouse: (1) a regression line, represented by the dashed line, showing the (estimated) relationship between time and demand and (2) a one-year-out estimate or forecast for future demand, represented by the shaded areas. Technically, the shaded area shows the 95 percent confidence intervals based on random samples. That is, there is a 95 percent chance that the shaded areas contain the true value of future demand.

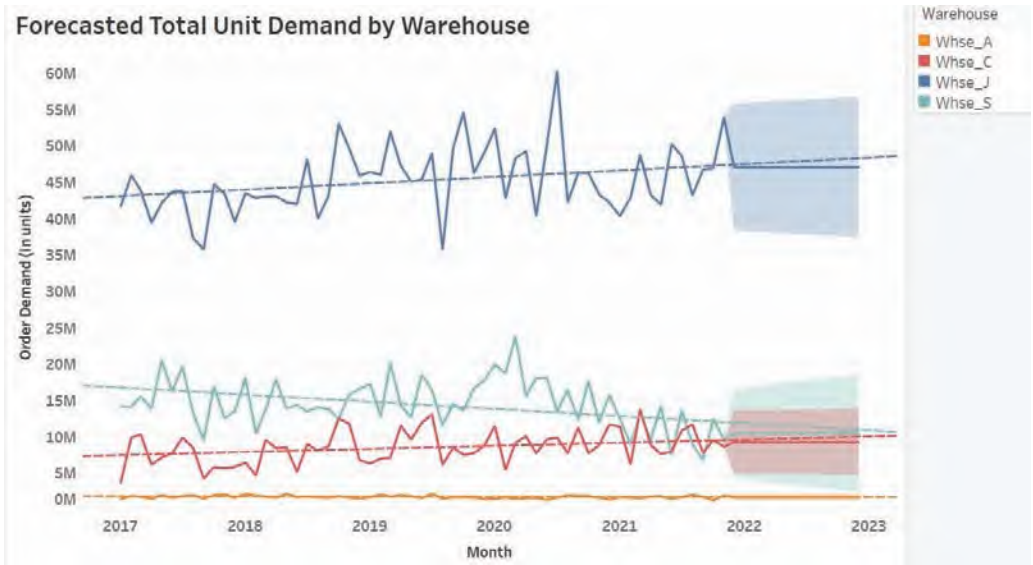


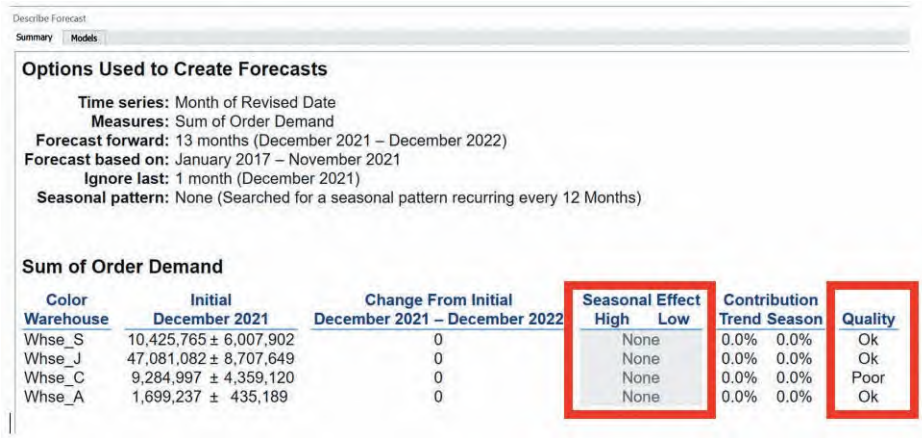
Exhibit 10.16
Forecasted Demand in Tableau

[22]These data were adapted from https://www.kaggle.com/felixzhao/productdemandforecasting, accessed on 7/20/2020.

The analysis shows that Warehouse J has much higher demand than the other warehouses, while Warehouse A has much lower demand than the other warehouses. In addition, demand at Warehouse J and Warehouse C seem to be trending slightly upwards while demand at Warehouse S appears to be falling. The company can use the predicted demand to help determine the appropriate head count needed at each warehouse in 2022.

Exhibit 10.17 presents a description of the forecast created by Tableau. No *seasonality,* a repeating pattern of variation over time, was detected for any of the warehouses, as indicated by "None" in the Seasonal Effect column. In addition, Tableau reports the Quality of the forecasts as "OK" except for Warehouse C, which is "Poor," indicating that the forecast has more error. As more data become available for analysis, forecasts should become more accurate, which should lead to "Good" forecast quality.

Exhibit 10.17
Forecasted Demand Description in Tableau



Describe Forecast

Summary | Models

Options Used to Create Forecasts

Time series: Month of Revised Date
Measures: Sum of Order Demand
Forecast forward: 13 months (December 2021 – December 2022)
Forecast based on: January 2017 – November 2021
Ignore last: 1 month (December 2021)
Seasonal pattern: None (Searched for a seasonal pattern recurring every 12 Months)

Sum of Order Demand

Color Warehouse	Initial December 2021	Change From Initial December 2021 – December 2022	Seasonal Effect High	 Low	Contribution Trend	 Season	Quality
Whse_S	10,425,765 ± 6,007,902	0	None		0.0%	0.0%	Ok
Whse_J	47,081,082 ± 8,707,649	0	None		0.0%	0.0%	Ok
Whse_C	9,284,997 ± 4,359,120	0	None		0.0%	0.0%	Poor
Whse_A	1,699,237 ± 435,189	0	None		0.0%	0.0%	Ok

LAB CONNECTION

LAB 10.5 provides step-by-step instructions on how to create a trend line and a forecast for supply-chain demand in Tableau.

PROGRESS CHECK

14. Which predictive analytics technique should be used to help hire new employees?
15. Which predictive analytics technique should be used to help develop profitable products?
16. Why does the size of the band around the forecast change when the confidence interval changes?

LO 10.7

Describe and provide examples of prescriptive analytics used to answer operations questions.

10.7 PRESCRIPTIVE OPERATIONS ANALYTICS

Prescriptive analytics answer the questions "What should we do, based on what we expect to happen?" and "How do we optimize our performance based on potential constraints?" While both predictive and prescriptive analytics make predictions about the future, prescriptive analytics recommends a course of action. Exhibit 10.18 identifies some prescriptive analytics techniques as well as examples of operations questions that each technique could answer.

		Example Operations Question		
Statistical Technique for Prescriptive Analytics	**Use**	**Human Resources**	**IT Operations**	**Supply Chain**
Cash Flow Analysis	Evaluates future cash flows for potential investments or expenditures; typically uses net present value (NPV) or internal rate of return (IRR)	Should a new training program be implemented?	Which computer system should be implemented?	How many delivery trucks should the company purchase?
Goal-Seek Analysis	A type of what-if analysis that determines which input value(s) are needed to achieve a desired output (that is, backward analysis)	How many employees are needed to staff a store during the holiday season?	How many IT support staff are needed to resolve all IT tickets within 30 minutes?	How many units need to be sold to break even?
Optimization	The process of making something as effective as possible, or at least better than it is now	What is the optimal number of employees to maximize profitability?	What is the lowest-cost method of processing transactions?	What is the optimal shipping route to minimize costs?
Scenario Analysis	A type of what-if analysis that examines the effect of changing multiple input variables at the same time on an outcome variable (typically, base/most likely case, worst case, and best cases)	What will employee turnover be?	How many potential security breaches will the company experience in the next year?	What is the optimal inventory policy to maintain a specified item fill rate (that is, the percent of customer orders satisfied by inventory on hand)?
Sensitivity Analysis	A type of what-if analysis that examines the effect of changing one input on an outcome variable	What will be employee turnover if a specific training program is implemented?	How many security breaches will there be next year if a new security program is installed?	What will be the fill rate if a specific inventory policy is implemented?

Exhibit 10.18
Prescriptive Analytics Techniques and Operations Examples

Optimization

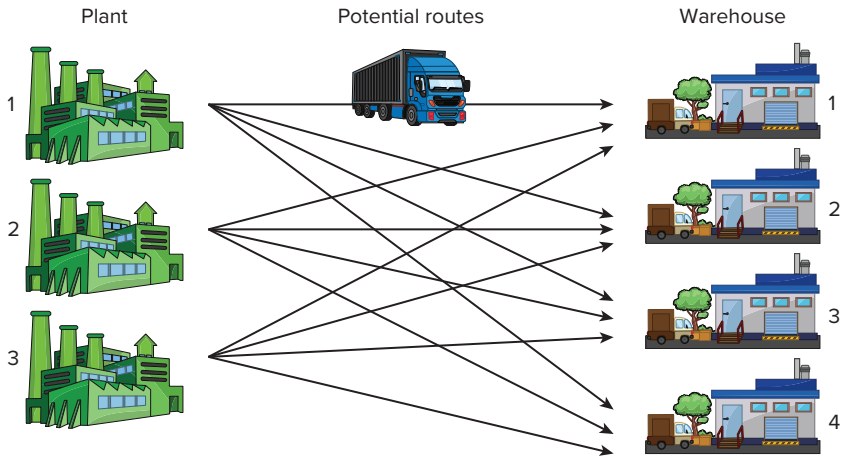
Optimization is the process of making something as effective as possible (or at least better than it is now). For example, companies typically want to maximize profit and minimize costs. In the supply chain, determining the best route to ship goods is important for controlling costs and delivering goods in a timely fashion, but it can be very complex. For example, **UPS** delivers over 21.9 million packages and documents a day using over 125,000 vehicles.[23] Across the United States, **Walmart** has 190 distribution centers[24] that support almost 4,800 U.S. stores and 600 **Sam's Clubs**.[25] Worldwide, Walmart maintains its own private fleet consisting of more than 6,500 tractors and 55,000 trailers, and it employs

[23]UPS, 2019, "UPS Fact Sheet." Available online at https://pressroom.ups.com/pressroom/ContentDetailsViewer.page?ConceptType=FactSheets&id=1426321563187-193. Accessed 7/21/2020.
[24]MWPVL International Inc., 2020, "Leadership in Global Supply Chain and Logistics Consulting." Available online at https://mwpvl.com/html/walmart.html. Accessed 7/21/2020.
[25]Walmart, 2020, "Location Facts." Available online at https://corporate.walmart.com/our-story/our-locations. Accessed 7/21/2020.

7,000 drivers who drive, on average, 100,000 miles a year to deliver goods in supersized trailers that can haul up to 40 percent more merchandise than traditional trailers.[26] To minimize wasted miles, both UPS and Walmart determine the best route for their drivers.

Can you imagine trying to identify the best routes for all of the UPS or Walmart trucks? We sure can't. So instead, let's look at a simplified example. Our goal is to fulfill demand at the lowest cost. Let's assume that a company manufactures one product in three manufacturing plants (Plant 1, 2, and 3), that supply four warehouses (Warehouse 1, 2, 3, 4). Exhibit 10.19 shows the potential shipping routes between the plants and warehouses.

Exhibit 10.19
Possible Shipping Routes



The per-unit cost to deliver from one of the plants to one of the warehouses varies. Each plant has a maximum capacity for manufacturing a product (3,500, 2,750, and 4,600, in that order). Each warehouse has its own customer demand for the product (2,000, 1,500, 5,250, and 2,100). The number of products that each plant can ship is less than or equal to its capacity. The number of products that each warehouse can receive must be greater than or equal to its demand. The goal is to minimize total shipping cost. Using the Solver Add-In for Excel, we can solve this problem. (See Appendix F for installation instructions.) The result is shown in Exhibit 10.20.

Based on the solution, Plant 1 should ship 1,500 units to Warehouse 2 and 2,000 units to Warehouse 4. Plant 2 should ship 2,650 units to Warehouse 3 and 100 units to Warehouse 4. Plant 3 should ship 2,000 units to Warehouse 1 and 2,600 units to Warehouse 3. The total cost for this shipping model is $14,599. The solution fulfills demand at all four warehouses at the lowest cost. Exhibit 10.21 visually shows the solution for our simplified example.

If you select Sensitivity when you run the Solver, Excel automatically generates a sensitivity report, which can help you understand the recommended shipping model as well as help identify potential places to try and reduce costs. Exhibit 10.22 shows the first four columns of the Variable Cells portion of the Report. If a Final Value equals 0 (meaning nothing was shipped between the specified plant and warehouse), then the Reduced Cost

[26]Walmart, 2021, "Logistics." Available online at https://walmartsupplychain.weebly.com/logistics.html. Accessed 7/08/2021.

Network Transportation Problem

Cost per unit delivery

	Warehouse 1	Warehouse 2	Warehouse 3	Warehouse 4
Plant 1	1.80	1.86	1.88	1.75
Plant 2	1.26	1.46	1.10	1.22
Plant 3	1.18	1.92	1.12	1.92

Recommend Shipping Model

	Warehouse 1	Warehouse 2	Warehouse 3	Warehouse 4	Ship		Capacity
Plant 1	0	1500	0	2000	3500	<=	3500
Plant 2	0	0	2650	100	2750	<=	2750
Plant 3	2000	0	2600	0	4600	<=	4600
							10850 Total
Received	2000	1500	5250	2100			
	>=	>=	>=	>=			
Demand	2000	1500	5250	2100	10850	Total	

TotalCost	$14,599.00	$ 2,360.00	$ 2,790.00	$ 5,827.00	$ 3,622.00

Exhibit 10.20
Example of a Simple Transportation Optimization Problem

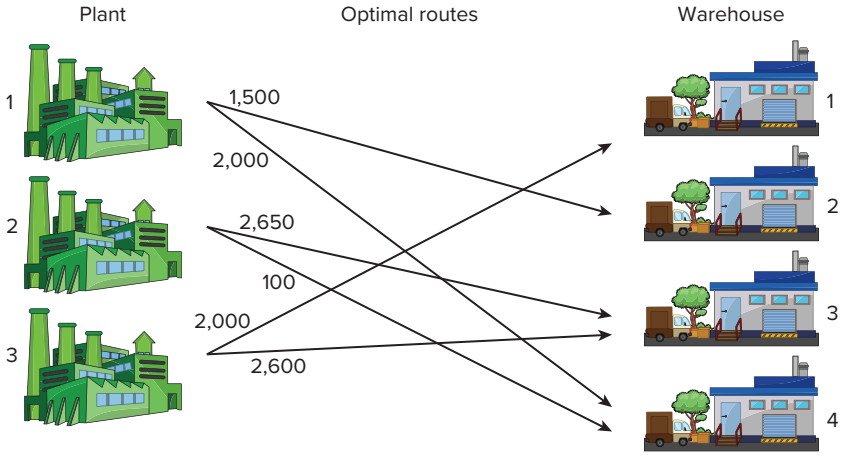


Exhibit 10.21
Optimal Solution to the Simple Transportation Optimization Problem

Variable Cells

Cell	Name	Final Value	Reduced Cost
$D$11	Plant 1 Warehouse 1	0	0.11
$E$11	Plant 1 Warehouse 2	1500	0
$F$11	Plant 1 Warehouse 3	0	0.25
$G$11	Plant 1 Warehouse 4	2000	0
$D$12	Plant 2 Warehouse 1	0	0.1
$E$12	Plant 2 Warehouse 2	0	0.13
$F$12	Plant 2 Warehouse 3	2650	0
$G$12	Plant 2 Warehouse 4	100	0
$D$13	Plant 3 Warehouse 1	2000	0
$E$13	Plant 3 Warehouse 2	0	0.57
$F$13	Plant 3 Warehouse 3	2600	0
$G$13	Plant 3 Warehouse 4	0	0.68

Exhibit 10.22
Excel Solver Sensitivity Report

field indicates how much the cost per unit must decrease before the specified plant would ship to the specified warehouse in the Name field. For example, for Plant 1 to ship products to Warehouse 1, the shipping cost must decrease by a little more than $0.11 per unit. For Plant 1 to ship products to Warehouse 3, shipping costs must drop by $0.25 per unit. The sensitivity analysis helps companies figure out the routes on which they should try and reduce costs in order to keep fulfilling demand while reducing total cost. Each time costs change, the Solver and sensitivity analysis should be re-run because the optimal routes and total cost may change.

LAB CONNECTION

LAB 10.6 provides step-by-step instructions on how to use Excel's Solver Add-In to minimize the cost of labor in a manufacturing environment.

PROGRESS CHECK

17. Which prescriptive analytics technique should be used to decide whether or not to buy an assembly robot?
18. What is optimization and why is it part of prescriptive analytics?

LO 10.8

Explain how to report the results of operations analytics.



DATA VISUALIZATION

10.8 REPORT THE RESULTS

As noted in Chapter 6, there are numerous ways to report the results of business analysis. This section focuses on operations dashboards. An **operations dashboard** is a visual management tool that helps track and monitor KPIs to improve operations effectiveness. Each company needs to decide not only which operations KPIs are important for success but also the best way to present them on the dashboard. Separate dashboards (with appropriate KPIs) are typically created for human resources, IT operations, and supply chain.

Exhibit 10.23 presents an example of an HR dashboard. It includes information about head count, employee costs, employee satisfaction, turnover, and absenteeism. Note that some of the KPIs, such as satisfaction, are shown at a point in time, while others, such as employee costs, are shown over time compared to prior years. This is just one of many potential HR dashboards. Other dashboards could depict employee salaries, performance, development progress, training costs, and demographics.

Exhibit 10.24 presents a sample IT operations dashboard for the IT Services Desk. It analyzes all tickets submitted to the IT Services Desk, showing ticket status, source or customer, category, age, and SLA (service-level agreement) violations by month. This dashboard is also interactive, allowing drill down. So, if you click on a graphic, additional information is revealed. For example, if you click on tickets that are older than six months, a window opens, identifying the ticket number, IT service employee assigned to the ticket, and ticket category. Managers can use this drill-down information to determine if a certain employee tends to have service tickets for a long time, or if a certain ticket category takes longer to resolve. This dashboard is just one of many potential IT operations dashboards. Other dashboards could depict IT network activity, project management, cybersecurity, IT costs, and user IT engagement.

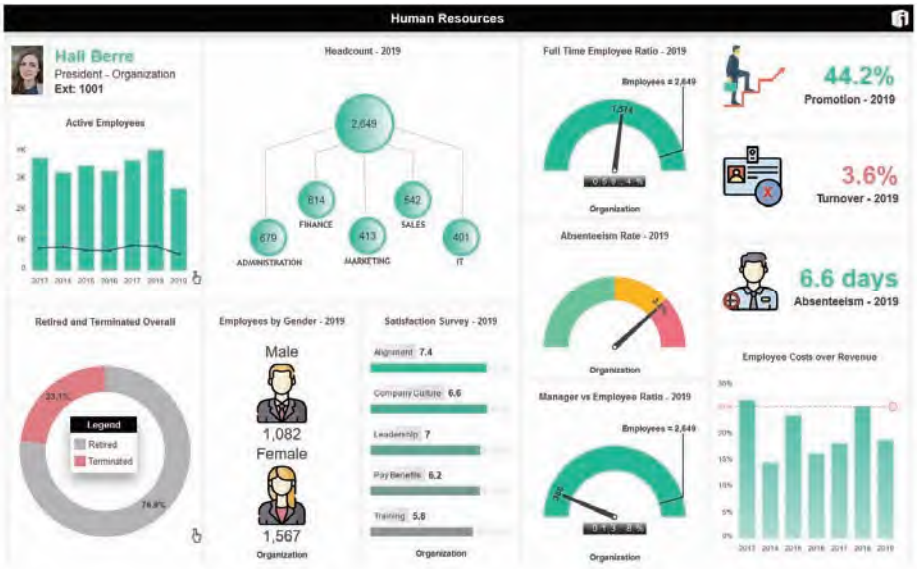


Exhibit 10.23
Sample HR Dashboard

Source: iDashboards, 2021, "Human Resources—HR Reporting." Available online at https://www.idashboards.com/dashboard-examples/human-resources-summary-dashboard/. Accessed 7/5/2021.



Exhibit 10.24
Sample IT Operations Dashboard

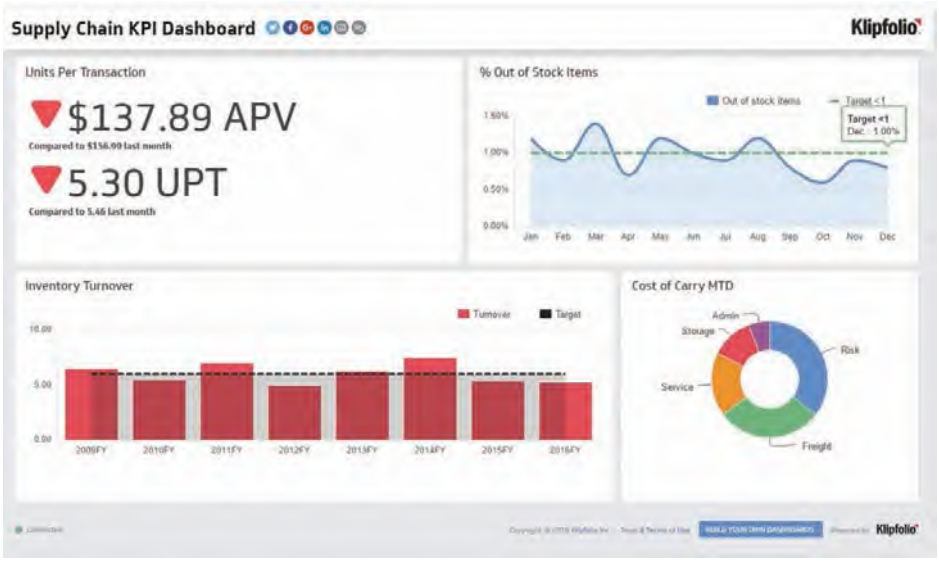
Source: iDashboards, 2020, "IT – Network Resource Monitoring." Available online at https://www.idashboards.com/dashboard-examples/it-dashboard-example-tech/. Accessed 7/5/2021.

Exhibit 10.25 shows a sample supply chain KPI dashboard. This dashboard includes the units per transaction (UPT, which were 5.3, a slight decrease from last month, as indicated by the downward-pointing chevron), average purchase value (APV, also down from last month), percent of out-of-stock items, inventory turnover, and cost of carrying

the inventory. This is just one example of a supply chain dashboard. Other supply chain dashboards could present information about procurement, transportation, and warehouse operations.

Exhibit 10.25
Sample Supply Chain Dashboard

Source: Klipfolio, 2020, "Supply Chain KPI Dashboard." Available online at https://www.klipfolio.com/resources/dashboard-examples/supply-chain/. Accessed 11/04/2021.



PROGRESS CHECK

19. Why is a dashboard a good tool for presenting the results of analytics?
20. What KPIs should **Walmart** include in its supply chain dashboard?

Chapter Summary

Operations analytics is used to answer operations questions. We specified the type of questions that can be addressed with both internal and external data across the three branches of operations (human resources, IT operations, and supply chain). Next, we discussed the data available to address operations questions. We explained the different types of analytics that can be performed, including descriptive, diagnostic, predictive, and prescriptive analytics. We concluded with examples of reporting the results of operations analytics with an operations dashboard.

Key Terms

algorithm (*707*) Computer-programmed instructions for how to combine data. For example, Total Costs = Fixed Costs + Variable Costs is an algorithm for how to calculate Total Costs.

control chart (*704*) A graph that plots data about a process in a time-ordered sequence.

fill rate (*697*) The percent of customer orders satisfied by inventory on hand.

human resources (HR) (*688*) The branch of operations that focuses on managing and developing a company's workforce.

human resource (HR) analytics (*690*) The processes and technologies that analyze people-based data to help managers make better HR decisions.

human resource management (HRM) system (*694*) A computer system that collects and stores human resource data in one place. It helps the company automate HR processes and comply with laws and regulations.

IT operations (*688*) A company's processes to manage its hardware, software, and IT support. The IT operations department is also responsible for collecting, storing, and managing data.

IT operations analytics (*690*) The processes and technologies that analyze data from various IT applications and IT infrastructure to help managers identify, predict, solve, and prevent IT-related issues.

key performance indicators (KPIs) (*700*) Numerical metrics used to gauge a company's performance.

operations (*688*) The processes that turn inputs into goods and services that companies sell.

operations dashboard (*714*) Visual management tool that helps track and monitor KPIs (key performance indicators) to improve operations. Separate dashboards are typically prepared for human resources, IT operations, and supply chain.

optimization (*711*) The process of making something as effective as possible (or at least better than it is now).

service-level agreement (SLA) (*695*) A contract that defines the level of service that a company expects from the technology vendor/provider as well as penalties if the agreed-on service levels are not attained.

Six Sigma (*704*) A set of techniques and tools used for continual process improvement almost to the point of zero defects.

supply chain (*689*) The branch of operations that is responsible for the production of a product or service and its distribution to the final consumer. The supply chain includes all of the processes necessary to make the product: obtaining raw materials, transforming those materials into the finished product, and transporting and distributing those products.

supply chain analytics (*690*) The processes and technologies that analyze supply chain data to help managers improve supply chain planning, efficiency, and effectiveness.

supply chain management (SCM) system (*695*) A computer system that collects and stores supply chain data in one place. It helps the company automate supply chain processes and comply with laws and regulations.

telemetry (*693*) The collection of data at remote points (possibly by a digital employee badge) and the automatic transmission of the data to monitoring equipment.

time series graph (*699*) A visual representation of a variable over time.

ANSWERS TO PROGRESS CHECKS

1. Both are branches of operations. IT operations refer to the processes that the company uses to manage its hardware, software, and IT support. Supply chain is responsible for the production of a product or service and its distribution to the final consumer. It includes all of the processes necessary to make the product, including obtaining raw materials, transforming raw materials into the finished product, and transporting and distributing the product.
2. Human resources and IT operations support the supply chain. Without the supply chain, a company will not have the need for human resources or IT support. (That is, no product to sell equals no company.)

3. Answers will vary but most likely depend on the type of company, industry, and the company's goals.
4. Potential HR questions addressable using analytics include "How has head count changed from last year?"; "Which new hires will perform best?"; and "How many employees does the company need during the holiday season?"
5. Potential supply chain questions addressable using predictive analytics include "What will demand be for a product?"; "What happens to costs if the company increases/decreases order quantities?"; and "Which machine will likely break down?"
6. Answers will vary. **Domino's Pizza** may examine historical data using descriptive analytics to answer the following questions: What was average system response time last year? How long on average did it take customers to place orders? What was average delivery time? How many orders were placed simultaneously? For diagnostic analytics, Domino's Pizza may ask: How has the average system response time changed over the past five years? How has average length of time to place an order changed over the past five years? For predictive analytics, Domino's Pizza may ask: What will system demand be for the upcoming Super Bowl? For prescriptive analytics, Domino's Pizza may ask: What is the best IT configuration to meet demand? What is the best data interface to reduce the average time to place an order?
7. Potential sources of internal data that can be used for operations analytics include data stored in human resource management systems, telemetry devices, computer logs, and supply chain management systems.
8. Potential sources of external data that can be used for operations analytics include **LinkedIn**, job websites, social media, weather, road conditions, and traffic patterns.
9. Answers will vary. **Domino's Pizza** may use internal sales data generated by its enterprise system as well as IT operations data to measure system performance. With respect to external data, Domino's Pizza may include measures of the teams' popularity, the health of the economy, and customers' discretionary income.
10. Days of payables is the number of days it takes for a company to pay its vendors. Days of inventory is the number of days a company holds inventory before it is sold. Days of receivables is the number of days after a sale until the company receives cash from its customers.
11. Answers will vary. **Domino's Pizza** may be interested in average system response time, average number of orders processed each minute, maximum number of customers ordering pizza at the same time, and maximum system response time.
12. The difference is that descriptive analytics would calculate a KPI at one point in time, while diagnostic analytics would calculate how a KPI changed over time.
13. Answers will vary. **Ford Motor Company** could use regressions and correlations to see if different variables/problems are related. Ford could also use pivot tables and summarizations to determine if certain vendors/suppliers have problems across Ford brands (such as Buick and Lincoln).
14. Classification can help identify potential new hires based on traits that have been shown to be important for company success.
15. Decision trees can help identify different options for developing products. Probabilities can help identify profitable products.
16. The size of the band around the forecast reflects the confidence interval associated with the forecast. For example, if you wanted to be only 10 percent confident in your forecast, then a very narrow confidence band would be used because then your forecast would have to fall within the band only 1 out of 10 times! In that case, your forecast would not be in that band 90 percent of the time. If you wanted to be 99 percent confident in your forecast, then a very wide confidence band would be used because your forecast would have to fall within the band 99 out of 100 times. The wider band covers more potential forecasts, increasing your odds of falling in the band.
17. Cash flow analysis is the proper prescriptive analytics technique because the company needs to evaluate future cash flows from investing in the assembly robot. Net present value (NPV) or internal rate of return (IRR) are appropriate techniques.

18. Optimization is the process of making something as effective as possible, or at least better than it is now. Optimization is a prescriptive analytics technique because it answers the question "How do we optimize our performance based on potential constraints?"
19. There are many reasons why a dashboard is a good tool for presenting the results of analytics: Much information is gathered in one place; the dashboard provides instant access to critical information; KPIs can be automatically evaluated against a target; and the dashboard summarizes a lot of data using easy-to-read visuals.
20. **Walmart** could include many potential KPIs in its dashboard, such as inventory turnover, average delivery times, top-selling products, top-performing vendors, and days of payables.

connect

Multiple-Choice Questions

1. (LO 10.1, 10.2) Which of the following is not part of operations?
   a. Advertising
   b. Integrating information technology
   c. Ordering raw materials
   d. Scheduling production
2. (LO 10.3) Which of the following collects data at remote points and transmits the data to monitoring equipment?
   a. Fill rate
   b. Operations dashboard
   c. Supply chain
   d. Telemetry
3. (LO 10.8) If an executive wanted to see more detailed information about a KPI on a dashboard, which feature would be helpful?
   a. Drill around
   b. Drill down
   c. Drill through
   d. Drill up
4. (LO 10.1, 10.3) Which of the following statements about supply chain management (SCM) systems is true?
   a. SCM systems can be linked to customer relationship management systems.
   b. SCM systems are usually integrated modules of an enterprise system.
   c. SCM systems include processes from procurement to delivering the goods to the customer.
   d. SCM systems can be linked to customer relationship management systems, are usually integrated modules of an enterprise system, and include processes from procurement to delivering the goods to the customer.
5. (LO 10.2, 10.4) "How many new employees were hired during the year?" is a question that can be answered by:
   a. descriptive analytics.
   b. diagnostic analytics.
   c. predictive analytics.
   d. prescriptive analytics.

6. (LO 10.2, 10.5) "Why has the number of minutes of downtime for our websites increased from last year?" is a question that can be answered by:
   a. descriptive analytics.
   b. diagnostic analytics.
   c. predictive analytics.
   d. prescriptive analytics.
7. (LO 10.2, 10.6) "How many active users will be using the enterprise system in three years?" is a question that can be answered by:
   a. descriptive analytics.
   b. diagnostic analytics.
   c. predictive analytics.
   d. prescriptive analytics.
8. (LO 10.6) This chapter presented some forecasts created in Tableau. How are forecasts differentiated from the actual data?
   a. Bolded lines are used for forecasts.
   b. Dashed lines are used for forecasts.
   c. Different colored lines are used for forecasts.
   d. Shading is used for forecasts.
9. (LO 10.2, 10.7) "Which employees need to go to training to improve their productivity?" is a question that can be answered by:
   a. descriptive analytics.
   b. diagnostic analytics.
   c. predictive analytics.
   d. prescriptive analytics.
10. (LO 10.7) To complete an optimization in Excel, which feature would you use?
    a. Data Analysis ToolPak
    b. Forecast sheet
    c. Solver Add-In
    d. What-if analysis
11. (LO 10.5) Companies want to retain good employees as long as possible because hiring and training new employees is expensive and time consuming. Which of the following techniques can help a company determine if the employee turnover rate is higher than the desired employee turnover rate, so that the company can investigate the causes of the high turnover rate and then make needed changes?
    a. Algorithm
    b. Control chart
    c. Optimization
    d. Telemetry
12. (LO 10.6) Assume that **Hertz** wants to understand the relationship between different variables, such as miles driven, car age, and annual maintenance costs. Which of the following techniques can help Hertz understand the relationship?
    a. Control chart
    b. Forecasting
    c. Regression
    d. Six Sigma
13. (LO 10.6) Assume that **Hertz** wants to estimate the annual maintenance costs for its fleet next year. Which of the following techniques can help Hertz make this prediction?
    a. Control chart
    b. Forecasting

c. Regression

d. Optimization

14. (LO 10.7) **Tesla** uses robots to make most of its automotive parts. Assume that the company is trying to decide if it should completely automate the production of its leather seats, which are currently made by hand. This change would require a huge investment in additional robotic technology. Which technique would help Tesla decide whether or not to make the investment?

    a. Cash flow analysis

    b. Goal-seek analysis

    c. Optimization

    d. Sensitivity analysis

15. (LO 10.7) Assume that **Publix**, the largest employee-owned supermarket chain in the United States, has calculated the net present value (NPV) for an investment in a new automated warehouse system. Which of the following statements is correct?

    a. Publix should invest in the system if the NPV is greater than or equal to zero.

    b. Publix should invest in the system if the NPV is negative.

    c. Publix should not invest in the system if the NPV is greater than or equal to zero.

Discussion Questions

1. (LO 10.1, 10.2) Describe the role of operations as well as the role of operations analytics in a company.
2. (LO 10.3) A human resource management (HRM) system can be set up to include information about a (potential) employee from the person's public social media activity, including likes/dislikes and tweets. Do you believe it is ethical for a company to collect this information? Why or why not?
3. (LO 10.3) How can companies use telemetry to gather information about employees? Do you believe that it is ethical for companies to use telemetry in this manner? Why or why not?
4. (LO 10.3, 10.4, 10.5, 10.6, 10.7) Describe how companies can analyze the digital footprint left by employees.
5. (LO 10.3, 10.4, 10.5) Assume that a company has standard operating hours of Monday through Friday, 8 A.M. to 5 P.M. What would be the concern if an employee is entering transactions at 3 A.M. on the weekend? How could a company detect this employee's activities?
6. (LO 10.3) Why would a company want to link its SCM system with its CRM system?
7. (LO1 0.4, 10.5) Let's examine the **Dell** computer order process. A customer places an online order for a computer and then pays for the order. After the order is placed and paid for, Dell orders the parts, vendors deliver the parts, the computer is assembled, and the computer is then shipped out to the customer. Dell pays the vendor when the invoice payment is due. How does this process affect the cash-to-cash cycle for Dell?
8. (LO 10.4, 10.5) Why would a company want to evaluate the IT operations KPI of unsolved tickets per IT service desk employee? How could a company use a KPI equal to the number of unsolved tickets divided by the number of solved tickets?
9. (LO 10.7) How could a company use a sensitivity report generated by the Solver Add-In to improve shipping operations?
10. (LO 10.8) Should companies have different dashboards for human resources, IT operations, and supply chain? Or should one overall dashboard be used, incorporating measures from all of the branches? Explain.

connect

Brief Exercises

1. (LO 10.1, 10.2) Match the branch of operations (Human Resources, IT Operations, Supply Chain) to the operations questions.

Operations Question	**Branch of Operations** (Human Resources, IT Operations, Supply Chain)
How many minutes of downtime did the company website have last month?	
How many employees have been with the company less than five years? Less than one year?	
What will product demand be in five years?	
What is the best shipping route to minimize cost and delivery time?	
Why has application performance changed?	
How many employees does the company need to work on a holiday?	
Which machine needs maintenance to prevent downtime?	

2. (LO 10.2, 10.4, 10.5, 10.6, 10.7) Match the analytics type (descriptive, diagnostic, predictive, or prescriptive) to the operations questions.

Operations Question	**Analytics Type** (Descriptive, Diagnostic, Predictive, or Prescriptive)
Why did the perfect order delivery rate change from last year?	
Which vendor is most likely to deliver on time?	
How many new machines should the company purchase?	
What is the median order delivery rate?	
Why did average delivery time increase?	
How many IT service desk employees are needed to resolve all IT tickets within 15 minutes?	
What is the internal promotion rate?	

3. (LO 10.3) Match the operations data to the branch of operations (Human Resources, IT Operations, Supply Chain).

Operations Data	**Branch of Operations** (Human Resources, IT Operations, Supply Chain)
Performance evaluations	
Network statistics	
Demand forecasts	
LinkedIn account	
Shipping routes	
Delivery times	
IT service tickets open and closed	

4. (LO 10.4, 10.5, 10.6, 10.7) Match the analytics type (descriptive, diagnostic, predictive, or prescriptive) and branch of operations (Human Resources, IT Operations, Supply Chain) to the analytics technique used.

Analytics Technique Used	**Analytics Type** (Descriptive, Diagnostic, Predictive, or Prescriptive)	**Branch of Operations** (Human Resources, IT Operations, Supply Chain)
Total number of employees		
Forecast to estimate product demand in two years		
Hypothesis test to see if server response time has increased		
Optimization to schedule assembly-line workers		
Scenario analysis to determine the optimal fill rate		
Regression to determine a likely server breakdown		
Count of employees by department		

5. (LO 10.4, 10.5, 10.6, 10.7) Match the analytics type (descriptive, diagnostic, predictive, or prescriptive) to the analytics technique used.

Analytics Technique Used	**Analytics Type** (Descriptive, Diagnostic, Predictive, or Prescriptive)
Optimization	
Forecasting	
Minimums	
Hypothesis testing	
Classification	
Goal seek	
Bar chart	

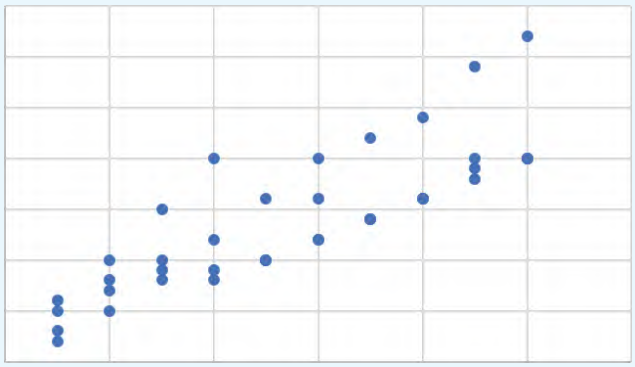
6. (LO 10.4, 10.5, 10.6, 10.7) Match the Excel menu option (Data Analysis or Solver) to the analytics technique used.

Analytics Technique Used	**Excel Menu Option** (Data Analysis or Solver)
Optimization	
t-test	
Regression	
Profit maximization	
Cost minimization	

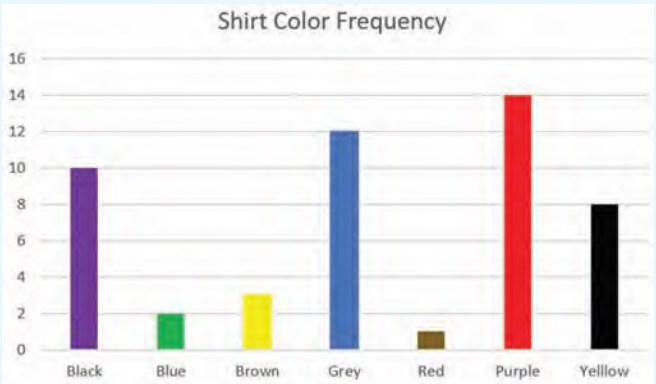
Problems

1. (LO 10.2, 10.4, 10.5, 10.6, 10.7) Companies may augment their SCM systems with a warehouse management system from software vendors such as **Oracle**, **SAP**, **InfoPlus**, and **Logiwa**. Describe specific types of warehouse management questions that can be answered using descriptive, diagnostic, predictive, and prescriptive analytics (one question for each). Describe the analytics techniques that could be used to address each of those questions.

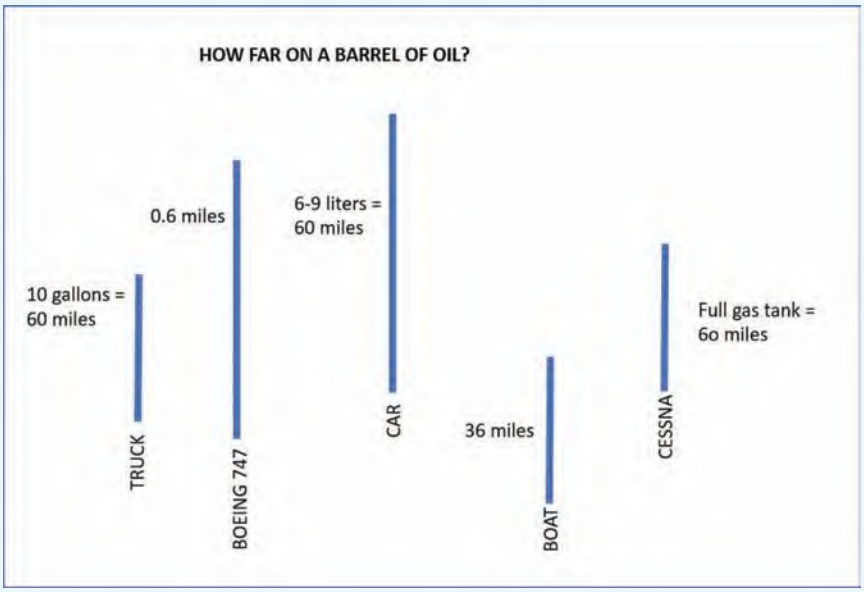
2. (LO 10.6) Business analysts can use regression to assess the relationship between variables and to predict future dependent-variable outcome values based on a given independent variable. For the following scenarios, identify the independent and dependent variables for a regression analysis.
   a. **UPS** wants to determine if weather affects on-time delivery.
   b. **PwC** wants to know if age is related to employee turnover.
   c. **Home Depot** wants to know the expected resolution time for a given level of IT service desk tickets.
3. (LO 10.5) A scatterplot shows relationships between variables. The following scatterplot graphs the relationship between number of users in the system and system response time. Correlation determines the extent to which variables are related to each other. Which variable is the X value? Which variable is the Y variable? How are these data points related?



4. (LO 10.1, 10.8) Data visualization should help companies understand how they are performing. The following data visualization was designed to answer this question: "How many thousands of shirts of each color were produced during the month?" How well does this data visualization answer this question, and how clearly does it present the results? Which branch of operations would this data visualization apply to? What type of analytics question is the data visualization answering?



5. (LO 10.8) Data visualizations should clearly present information. Evaluate the clarity of the following data visualization, which answers the question: "How far can different modes of transportation go on one gallon of gas?"



6. (LO 10.6) The Regression option of the Excel Analysis ToolPak was used to determine the relationship between total cost and production units. The results are shown below. What is the regression equation? How well does the model fit the data? What is the total cost for 1,200 units?

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.93
R Square	0.87
Adjusted R Square	0.86
Standard Error	6163.22
Observations	12

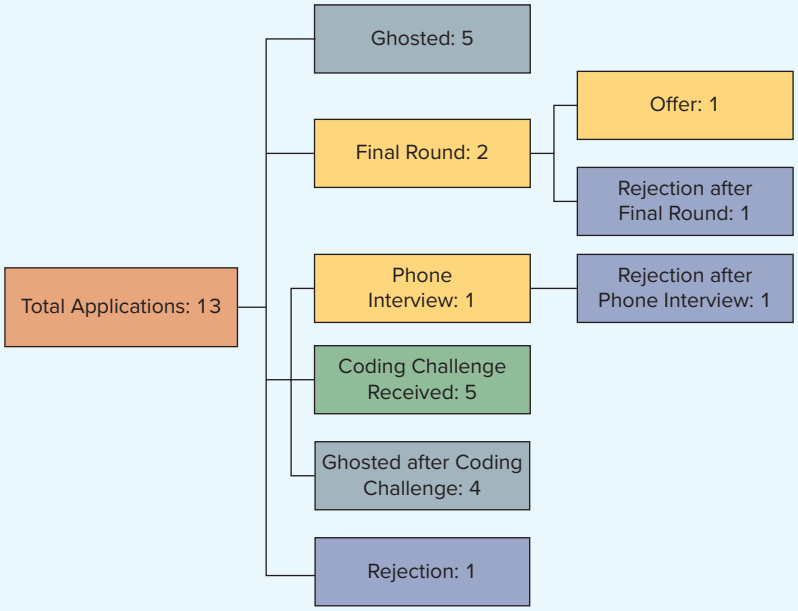
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2585890234.80	2585890234.80	68.08	8.97597E-06
Residual	10	379852265.20	37985226.52		
Total	11	2965742500.00			

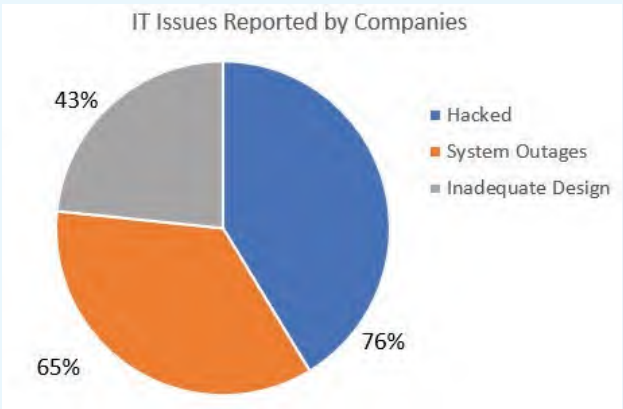
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	24287.01	10548.62	2.30	0.04	783.21	47790.80
Production Units	76.62	9.29	8.25	0.00	55.93	97.32

7. (LO 10.4, 10.5, 10.6, 10.7, 10.8) Select a branch of operations analytics. Decide which KPIs you would like to see on a dashboard. Identify a target audience and design an appropriate dashboard layout. Provide a rationale for your selected KPIs and dashboard design.

8. (LO 10.8) Evaluate the design of this data visualization for the hiring process.



9. (LO 10.4, 10.8) A chief information officer (CIO) is the corporate executive in charge of the strategy and implementation of the company's IT. In other words, the CIO is in charge of all IT personnel, processes, and technologies. Assume that a survey of CIOs was conducted. The survey asked the CIOs if their company had the following issues: (1) system was hacked, (2) system experienced outages, and (3) systems were inadequate to support the company's processes. Each CIO could select none, one, two, or all three issues. The following graphic of the results was created. Evaluate the design and create your own data visualization for the results.



10. (LO 10.7) **Apron Outfitters** is a manufacturer of aprons, jackets, and vests for a variety of occupations, including golf caddies, chefs, and carpenters. Assume that Apron Outfitters needs to schedule shipments of aprons from its three manufacturing plants to its four warehouses. The company has generated the following model using Excel's Solver Add-In. Using the model results, answer the following questions:
    a. What is the goal of the model?
    b. What is the total cost of the recommended shipping model?
    c. How does total capacity compare to demand?
    d. How many units need to be shipped from Plant 1 to Warehouse 3?
    e. How much does cost need to change for Plant 2 to ship to Warehouse 3?

Network Transportation Problem										
Cost per unit delivery										
			Warehouse 1	Warehouse 2	Warehouse 3	Warehouse 4				
		Plant 1	2.69	3.40	3.09	2.45				
		Plant 2	2.44	3.12	5.08	5.67				
		Plant 3	2.02	4.32	4.51	3.21				
	Recommend Shipping Model									
			Warehouse 1	Warehouse 2	Warehouse 3	Warehouse 4	Ship		Capacity	
		Plant 1	450	0	689	250	1389	<=	2000	
		Plant 2	200	1300	0	0	1500	<=	1500	
		Plant 3	600	0	0	0	600	<=	600	
									4100	Total
		Received	1250	1300	689	250				
			>=	>=	>=	>=				
		Demand	1250	1300	689	250	3489	Total		
	TotalCost	$ 9,704.42	$ 2,909.84	$ 4,051.39	$ 2,129.50	$ 613.69				

Microsoft Excel 16.0 Sensitivity Report
Worksheet: [Problems.xlsx]Problem 10
Report Created: 2/16/2022 12:30:13 PM

Variable Cells

Cell	Name	Final Value	Reduced Cost
$D$11	Plant 1 Warehouse 1	450	0.00
$E$11	Plant 1 Warehouse 2	0	0.03
$F$11	Plant 1 Warehouse 3	689	0.00
$G$11	Plant 1 Warehouse 4	250	0.00
$D$12	Plant 2 Warehouse 1	200	0.00
$E$12	Plant 2 Warehouse 2	1300	0.00
$F$12	Plant 2 Warehouse 3	0	2.24
$G$12	Plant 2 Warehouse 4	0	3.46
$D$13	Plant 3 Warehouse 1	600	0.00
$E$13	Plant 3 Warehouse 2	0	1.62
$F$13	Plant 3 Warehouse 3	0	2.08
$G$13	Plant 3 Warehouse 4	0	1.42

Data Visualization Exercises



DATA VISUALIZATION

1. Keeping track of employee overtime hours is extremely important to understanding staffing, employee morale and turnover, and labor costs. A company created the following non-interactive, hand-typed data visualization to document overtime by employees. It appears that column B is reporting the number of employees who worked overtime hours, and the number in rows 2–13 corresponds to a specific month. For example, 1 employee worked a total of 20 overtime hours in January; 2 employees worked 36 hours of overtime in February; and 3 employees worked a total of 76 hours of overtime in March.



Is there a better visualization to show total overtime hours for each month? Create an interactive visualization that makes it easy to determine the number of employees working overtime during a month and the total number of overtime hours for each month. What additional data would help to enhance the data visualization even more?

CHAPTER 10 LABS

LAB 10.1 **Excel:** Descriptive Analytics/Supply Chain: Calculating KPIs for Walmart and Amazon (or Bed, Bath, and Beyond and Target)

LAB 10.2 **Tableau, Power BI:** Descriptive Analytics/Human Resources: Analyzing Employee Turnover Using a Bar Chart

LAB 10.3 **Excel:** Diagnostic Analytics/Supply Chain: Analyzing KPI Differences Between Walmart and Amazon (or Bed, Bath, and Beyond and Target)

LAB 10.4 **Tableau:** Diagnostic Analytics/IT Operations: Analyzing Website Response Time Using a Control Chart

LAB 10.5 **Tableau:** Predictive Analytics/Supply Chain: Forecasting Unit Demand

LAB 10.6 **Excel:** Prescriptive Analytics/Human Resources: Optimizing Employee Assignment

Lab 10.1 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Descriptive Analytics/Supply Chain: Calculating KPIs for Walmart and Amazon

Keywords

Descriptive Analytics, KPIs, Excel Analysis ToolPak, Supply Chain

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" Exhibit 10.8 presented supply chain key performance indicators (KPIs) for the entire retail and wholesale trade industry. In this lab, you will calculate KPIs and then use the built-in functionality of Excel's Analysis ToolPak to generate descriptive analytics for supply chain data for two specific retail companies, **Amazon** (ticker: AMZN) and **Walmart** (ticker: WMT), for the period 1999–2021.

Required

1. Calculate the following supply chain KPIs: cash-to-cash cycle, asset turnover, inventory-to-sales ratio, and inventory turnover.
2. Calculate descriptive analytics for the KPIs.
3. Create a box-and-whisker plot for inventory turnover.

Specify the Question

What are selected supply chain KPIs for **Amazon** and **Walmart?**

Obtain the Data

This lab uses the data in Lab 10.1 Data.xlsx. The data dictionary for this file follows. The data include annual financial information for the years 1999–2021.[27]

[27]Wharton Research Data Services (WRDS), "Compustat North American Fundamentals–Annual." https://wrds-www.wharton.upenn.edu/ 3/13/2022.

Data Dictionary

Year:	Data (fiscal) year
Ticker:	Stock exchange ticker for company
Company:	Company name (**Amazon** or **Walmart**)
AccountsPayable:	Total trade accounts payable due to vendors at the end of the fiscal year
AccountsReceivable:	Total accounts receivable due from customers at the end of the fiscal year
Average Assets:	Average total assets at the end of the last two fiscal years
CostofGoodsSold:	Total cost of items sold for the fiscal year
SalesRevenue:	Total sales revenue for the fiscal year
AverageInventory:	Average total inventory at the end of the last two fiscal years

Analyze the Data

Analysis Task 1: Install the Analysis ToolPak.

1. The first step is to install the Analysis ToolPak in Excel if it is not already installed. Textbook Appendix E provides instructions for installing the ToolPak.

Analysis Task 2: Calculate KPIs.

Exhibit 10.8 presented several KPIs for the retail and wholesale trade industry. Below are formulas to calculate those KPIs.

(1) Cash-to-cash cycle

$$= \text{Days of Inventory} + \text{Days of Receivables} - \text{Days of Payables}$$

$$= \frac{365 * \text{AverageInventory}}{\text{CostofGoodsSold}} + \frac{365 * \text{AccountsReceivable}}{\text{SalesRevenue}} - \frac{365 * \text{AccountsPayable}}{\text{CostofGoodsSold}}$$

(2) Asset turnover = SalesRevenue/Average Assets
(3) Inventory-to-sales ratio = Average Inventory/SalesRevenue
(4) Inventory turnover = CostofGoodsSold/AverageInventory

1. Type the following titles (do not include quotation marks):

   Cell J1: "Days of Payables"
   Cell K1: "Days of Inventory"
   Cell L1: "Days of Receivables"
   Cell M1: "Cash-to-Cash Cycle"
   Cell N1: "Asset Turnover"
   Cell O1: "Inventory-to-Sales"
   Cell P1: "Inventory Turnover"

2. Type the following formulas in the designated cells (do not type the quotation marks):

   Days of Payables–Cell J2: "=D2/(G2/365)"
   Days of Inventory–Cell K2: "=(365*I2)/G2"
   Days of Receivables–Cell L2: "=(365*E2)/H2"
   Cash-to-Cash Cycle–Cell M2: "=K2+L2−J2"
   Asset Turnover–Cell N2: "=H2/F2"
   Inventory-to-Sales–Cell O2: "=I2/H2"
   Inventory Turnover–Cell P2: "=G2/I2"

3. Copy the formulas down the column.
4. Round the answers to two decimal places. (Use Decrease Decimal in the Home tab.) Lab Exhibit 10.1.1 presents the first several rows of the spreadsheet.

Lab Exhibit 10.1.1
Sample KPIs for Walmart

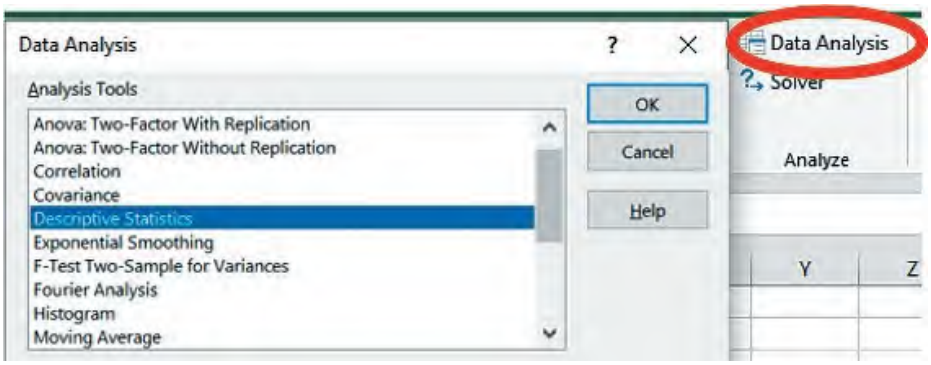
Days of Payables	Days of Inventory	Days of Receivables	Cash-to-Cash Cycle	Asset Turnover	Inventory-to-Sales	Inventory Turnover
37.47	52.70	2.96	18.19	2.75	0.11	6.93
37.25	50.89	3.36	17.00	2.59	0.11	7.17
33.76	47.61	3.34	17.20	2.70	0.10	7.67
33.15	45.93	3.14	15.92	2.75	0.10	7.95
36.14	48.14	1.78	13.78	2.58	0.10	7.58
36.71	47.48	2.19	12.96	2.54	0.10	7.69

Analysis Task 3: Calculate Descriptive Statistics Using the Excel Analysis ToolPak.

We are now ready to calculate descriptive statistics for the seven KPIs we just calculated.

1. Select **Data > Analyze > Data Analysis**. Select Descriptive Statistics, as shown in Lab Exhibit 10.1.2. Click OK.

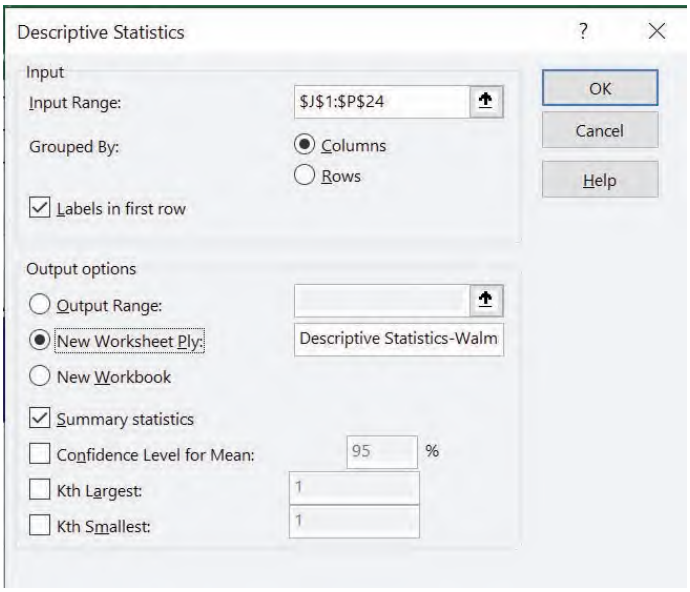
Lab Exhibit 10.1.2



2. Descriptive Statistics works only on numeric fields, so all of the KPIs are appropriate for the calculations. Because Descriptive Statistics calculates the statistics for an entire column, we need to manually split the Walmart and Amazon data. Let's calculate the descriptive statistics for Walmart first. In the Descriptive Statistics dialog box, enter the Input Range as "$J$1:$P$24". Accept the Group By Columns default. Check the box by Labels in the First Row. Select New Worksheet Ply: to have the results automatically placed in a new worksheet tab.[28] Enter "Descriptive Statistics–Walmart" in the field by New Worksheet Ply to automatically name the new worksheet. Check the box by Summary statistics. The Descriptive Statistics dialog box should look like Lab Exhibit 10.1.3. Click OK.

[28]Note: To put the descriptive statistics on the same worksheet, select Output Range and identify the place you want to show the results. To have the results placed in a completely new workbook or file without the Data sheet, select New Workbook.

Lab Exhibit 10.1.3



Excel should automatically generate the descriptive statistics. Format the statistics to two decimal places. Lab Exhibit 10.1.4 shows the output for Walmart's Days of Payables. Note the Average Days of Payables over the last 23 years (Count) was 40 days with a minimum of 33 days (Minimum) to a maximum of 48 days (Maximum). Thus, the Range is 48 – 33 = 15 days. Mode represents the value that occurs most frequently, but in this case there are no repeated observations, as noted by #N/A. Median represents the midpoint. Thus, half of the Days of Payables are above 40 days and half are below 40 days.

Standard Deviation is the square root of the Sample Variance and measures how far numbers are spread out from the average value. A higher Standard Deviation indicates that the data are more dispersed. For this set of data, one standard deviation from the mean (for Days of Payables) is between 35.90 days (39.87 – 3.97) and 43.84 days (39.29 + 3.57).

Kurtosis and Skewness describe the shape of the distribution, helping determine the likelihood of an event falling in a tail of a probability distribution. Kurtosis measures the thickness of the tails of the distribution. A kurtosis value of −0.30 indicates that the distribution has flat tails (versus the normal distribution, which has a kurtosis value of 3). Skewness equals 0.40, indicating that the distribution is slightly positively (right) skewed (versus normal distribution of 0).

Lab Exhibit 10.1.4

Walmart's Payable Days, 1999–2021

Days of Payables	
Mean	39.87
Standard Error	0.83
Median	39.55
Mode	#N/A
Standard Deviation	3.97
Sample Variance	15.77
Kurtosis	-0.30
Skewness	0.40
Range	15.07
Minimum	33.15
Maximum	48.21
Sum	917.04
Count	23.00

3. Repeat the steps above for Amazon. Type "Descriptive Statistics–Amazon" in the field by New Worksheet Ply to automatically name the new worksheet.



DATA VISUALIZATION

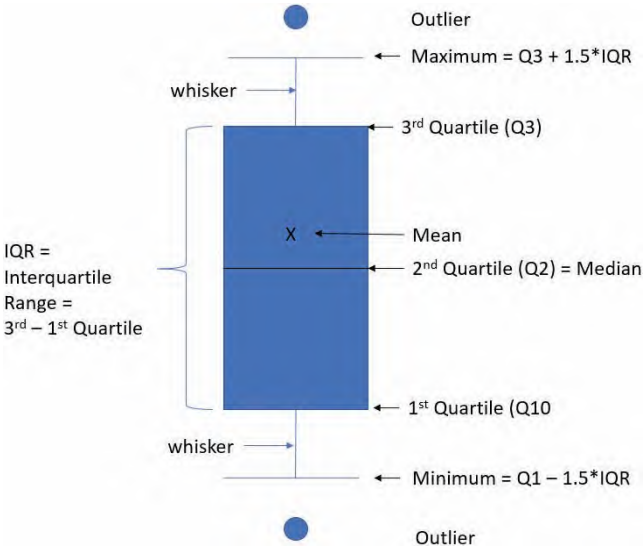
Quartile: Quartiles divide the data into four equal parts. The lower quartile contains values below the 25th percentile of the data, while the upper quartile contains the values above the 75th percentile of the data.

Report the Results: Create a Box-and-Whisker Plot

You have just (1) calculated select supply chain KPIs for retail giants Amazon and Walmart for the years 1999–2021 and (2) created descriptive statistics for the two companies. Financial analysts and other users of the financial statements can use the descriptive statistics to compare companies.

One way to communicate your descriptive statistics with an audience is through a box-and-whisker plot, which communicates variation in data. It shows the mean, median, and upper and lower quartiles of data, as well as any outliers. Lab Exhibit 10.1.5 presents the components of a box-and-whisker plot. The median is the point at which half of the data are below and half of the data are above the point as indicated by the 2nd quartile point (Q2). The 1st/3rd quartiles are the middle values of the lower/upper halves of the data. An outlier is outside the whiskers of the data plot.

Lab Exhibit 10.1.5



Let's create a box-and-whisker plot for Inventory Turnover, using the following steps.

1. Copy and paste the Company name column and Inventory Turnover column to a new Excel sheet. Your data should look like Lab Exhibit 10.1.6, with the Amazon data right below the Walmart data.

Lab Exhibit 10.1.6

	Company	Inventory Turnover
2	WALMART INC	6.93
3	WALMART INC	7.17
4	WALMART INC	7.67
5	WALMART INC	7.95
6	WALMART INC	7.58

2. Highlight all of the data in columns A and B.
3. From the menu, select **Insert > Charts > Insert Statistic Chart > Box and Whisker.** (*Note:* Your version of Excel may not include "Insert Statistic Chart.")

4. Change Chart Title to "Inventory Turnover Variation".
5. To add a Y-axis title, select **Chart Tools > Design > Add Chart Element > Axis Titles > Primary Vertical**. Excel adds "Axis Title" to the left of the Y-axis. Highlight the Axis Title and type "Inventory Turnover Ratio".

The resulting box-and-whisker chart will be as shown in Lab Exhibit 10.1.7. The plot clearly shows that while Amazon has a higher mean and higher median inventory turnover ratio, Walmart has a controlled inventory process with a very stable inventory turnover ratio. A box-and-whisker plot can be created for any of the KPIs used in this lab.

Lab Exhibit 10.1.7



Assessment

1. Take a screenshot of the completed descriptive statistics tables for Walmart and Amazon, copy it into a Word document named "Lab 10.1 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the box-and-whisker plot, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Lab 10.1 Excel Submission.docx via Connect, if assigned.

Alternate Lab 10.1 Excel

Descriptive Analytics/Supply Chain: Calculating KPIs for Bed, Bath, and Beyond and Target

Apply the same steps in Lab 10.1 to the Alt Lab 10.1 Data.xlsx data set.

Specify the Question

What are selected supply chain KPIs for **Bed, Bath, and Beyond** and **Target**?

Required

1. Calculate the following supply chain KPIs: cash-to-cash cycle, asset turnover, inventory-to-sales ratio, and inventory turnover.
2. Calculate descriptive analytics for the KPIs.
3. Create a box-and-whisker plot for inventory turnover.

Assessment

1. Take a screenshot of the completed descriptive statistics tables for Bed, Bath, and Beyond and Target, copy it into a Word document named "Alt Lab 10.1 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the box-and-whisker plot, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Alt Lab 10.1 Excel Submission.docx via Connect, if assigned.

Lab 10.2 Tableau

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Descriptive Analytics/Human Resources: Analyzing Employee Turnover Using a Bar Chart

Keywords

Descriptive Analytics, Bar Charts, Human Resources

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" One of the data visualizations used in descriptive analytics is a bar chart, which is useful for showing the distribution of data or performing a comparison of subgroups, or categories, of the data. Business analysts use bar charts to show which categories are the highest or lowest along some dimension. Bar charts can also show how the categories compare.

One of the roles of the human resources (HR) department is to monitor employee turnover—that is, the number of employees who leave the company. Finding and hiring employees takes time and money, so companies want to retain employees, especially those who perform at high levels. HR departments often monitor employee turnover by department (or some other dimension) to pinpoint departments (or policies) that may need changes to help the company retain employees. In this lab, you will use Tableau to generate two bar charts identifying (1) which department had the highest turnover, or the greatest number of employees who quit, and (2) the salary levels of employees who quit.

Required

1. Create a bar chart breaking down employee turnover by department.
2. Create a bar chart showing the salary levels for employees who quit.
3. Create a combined bar chart for employee turnover by department and salary.

Specify the Question

1. Which department had the highest number of employees quit?
2. What was the breakdown of employees quitting by salary level?

Obtain the Data

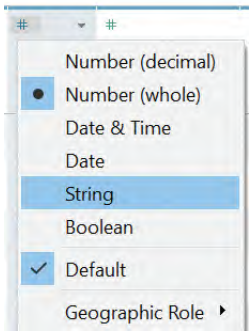
This lab uses the data in Lab 10.2 Data.xlsx. The HR data are from Kaggle.com and based on a real multinational company that we will call MNC Inc. The data dictionary for this file is below.

Data Dictionary

EMPID:	Unique employee ID
Satisfaction:	Satisfaction level of the employee (0–1: 0 = not satisfied, 1 = very satisfied)
Projects:	Number of projects that an employee is currently assigned to
MonthlyHours:	Average hours worked by the employee in the last three months
CommuteTime:	Monthly commute time to office
Quit:	1 = employee quit, 0 = employee is still employed by MNC
Promoted:	1 = employee promoted within the past five years, 0 = no promotion in past five years
Department:	Department employee works in
Salary:	Relative salary paid to employee (low, medium, high)

Obtain Task 1: Load the Data into Tableau.

1. Open Tableau.
2. Under Connect, select Microsoft Excel.
3. In the Open window, browse to the Lab 10.2 Data.xlsx file. Click Open. Notice that Tableau automatically separated the field names to different words when a capital letter was used in the Excel column headings (so MonthlyHours becomes Monthly Hours). Also notice that Tableau *properly* defined Employee ID as a string (see the Abc in the upper-left corner above Emp ID) because Emp ID is an identifier/dimension, not a number used in computations. However, Tableau defined Quit as a number (see the symbol # in the upper-left corner above Quit). Because Quit indicates a category of employees (Quit, Not Quit), not a number used in computations, change Quit to a string. To do this, click on the # above Quit and select String, as shown in Lab Exhibit 10.2T.1. Quit should now be identified as an "Abc" type of data. Repeat the steps to set Promoted to string data. The data are now ready to use.



Lab Exhibit 10.2T.1

Tableau Software

Analyze the Data

Analysis Task 1: Create the First Bar Chart: Number of Employees Leaving by Department.

In the first bar chart, we will determine the number of employees who quit in each department.

1. Click on Sheet 1.
2. Drag Department from the Data pane to the Columns shelf.
3. Notice that even though we changed the type of Quit to string (Abc), Tableau has created a measure that counts Quit. Drag Quit(Count) from the Data pane to the Rows shelf. CNT(Quit) should show in the Rows shelf counting the number of Quit values in the data. A bar chart with Department on the X-axis and Count of Quit on the Y-axis should be created, as shown in Lab Exhibit 10.2T.2.



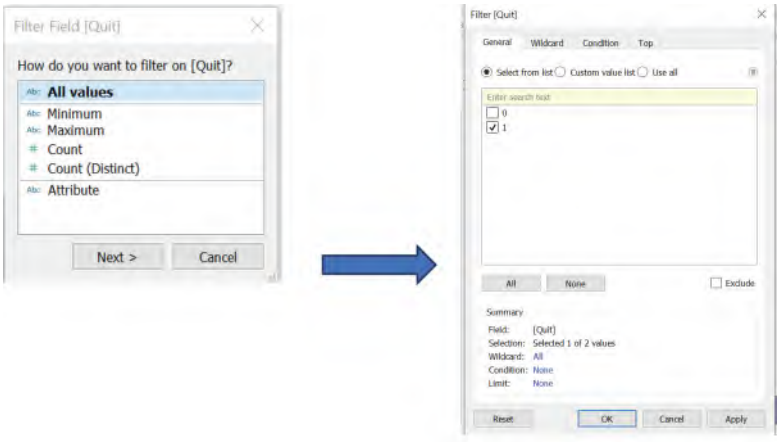
Lab Exhibit 10.2T.2

Tableau Software

4. Currently, the visualization shows all of the values of Quit, so it counts both 0s and 1s. We want only the employees leaving the company to show, or Quit = 1. We need to create a Filter to do this. To create a Filter, drag Quit(Count) to the Filters shelf. The Filter Field [Quit] dialog box appears. Select All values because you need to specify that you want only values equal to 1. Click Next >. In the Filter [Quit] dialog box, check 1. Click OK. Lab Exhibit 10.2T.3 shows what the dialog boxes should look like.

Lab Exhibit 10.2T.3

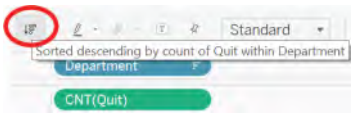
Tableau Software



5. Let's make the visualization easier to read. Sort the columns from highest to lowest using the "Sorted descending by count of Quit within Department" button at the top, as shown in Lab Exhibit 10.2T.4.

Lab Exhibit 10.2T.4

Tableau Software



6. To make the bar chart more precise, let's label each column with the exact number of employees quitting. Drag Quit(Count) from the Data pane to Label (T Label) in the Marks card. The bar chart should now clearly show the total number of employees quitting in each department.
7. Let's rename the Y-axis. Double-click on Count of Quit. In the Edit Axis [Count of Quit] dialog box, change the Title to "Number of Employees who Quit". Click the x at the top to close the dialog box.
8. Before moving to the next bar chart, let's name the data visualization. Double-click on Sheet 1 tab and type "Number of Employees who Quit by Department". Click Enter.

Analysis Task 2: Create the Second Bar Chart: Salary Level of Employees Quitting.

In the second bar chart, we will break down the number of employees who quit by salary level (low, medium, high).

1. Click on Sheet 2.
2. Drag Salary from the Data pane to the Columns shelf.
3. Drag Quit(Count) from the Data pane to the Rows shelf.
4. Sort the columns from highest to lowest using the same sort button as in Analysis Task 2.
5. Create a Filter for Quit(Count) to include only employees who quit (Quit = 1).

6. Label the columns with the precise number following the procedure outlined in Analysis Task 2.
7. Change the Y-axis label to "Number of Employees who Quit".
8. Let's now name the data visualization. Double-click on the Sheet 2 tab and type "Number of Employees who Quit by Salary Level". Click Enter.

Report the Results: Create a Combined Bar Chart



You have just calculated (1) the number of employees who quit in each department and (2) the number of employees who quit in each salary level. Management was not very specific about how it wanted the results reported. So, to help report the results to management, let's create a combined visualization of this information. Follow these steps:

1. Create a copy of the Numbers of Employees who Quit by Department tab. (Right-click on the tab and select Duplicate.)
2. Label it "Percent of Employees who Quit by Department and Salary Level". (Right-click on the tab and select Rename.)
3. Drag Salary to Color in the Marks card.
4. Drag CNT(Quit) to Label in the Marks card. Lab Exhibit 10.2T.5 shows the visualization. The data visualization clearly shows that 324 of the employees who quit were earning a low salary and worked for the Sales department.



Lab Exhibit 10.2T.5

Tableau Software

5. If we want to see the percentage of the total number of employees who quit, we click the down arrow by CNT(Quit) in the Marks card and select **Quick Table Calculation > Percent of Total.** The numbers are now percentages. As Lab Exhibit 10.2T.6 shows, 324 represents 16.2 percent of the total employees who quit. Although this visualization provides a great deal of information in one place, the percentages with decimals are overwhelming, as seen in Lab Exhibit 10.2T.6.

Lab Exhibit 10.2T.6

Tableau Software



6. To solve this issue, let's remove the decimals from the percentages. Click the down arrow by CNT(Quit) in the Marks card and select Format. The Format % of TOTAL CNT(Quit) window appears on the left. In the Default section by Numbers, click the drop-down arrow, select Percentage, and set Decimals equal to 0. Click somewhere out of the Format window. The percentages should now completely fit in each box, as shown in Lab Exhibit 10.2T.7. This combined visualization provides a great deal of helpful information to the management of MNC Inc.

Lab Exhibit 10.2T.7

Tableau Software



Assessment

1. Take a screenshot of the bar chart showing the number of employees who quit by department, copy it into a Word document named "Lab 10.2 Tableau Submission.docx", and label the screenshot Submission 1. Take a screenshot of the completed stacked bar chart for number of employees who quit by salary level, paste it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the completed stacked bar chart for the number of employees who quit by department and salary level, paste it into the same Word document, and label the screenshot Submission 3.
2. Answer the questions in Connect and upload your Lab 10.2 Tableau Submission.docx via Connect, if assigned.

Lab 10.2 Power BI

Descriptive Analytics/Human Resources: Analyzing Employee Turnover Using a Bar Chart

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Descriptive Analytics, Bar Charts, Human Resources

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Descriptive analytics answers the questions "What happened?" and "What is happening?" One of the data visualizations used in descriptive analytics is a bar chart, which is useful for showing the distribution of data or performing a comparison of subgroups, or categories, of the data. Business analysts use bar charts to show which categories are the highest or lowest along some dimension. Bar charts can also show how the categories compare.

One of the roles of the human resources (HR) department is to monitor employee turnover—that is, the number of employees who leave the company. Finding and hiring employees takes time and money, so companies want to retain employees, especially those who perform at high levels. HR departments often monitor employee turnover by department (or some other dimension) to pinpoint departments (or policies) that may need changes to help the company retain employees. In this lab, you will use Power BI to generate two bar charts identifying (1) which department had the highest turnover, or the greatest number of employees who quit, and (2) the salary levels of employees who quit.

Required

1. Create a bar chart breaking down employee turnover by department.
2. Create a bar chart showing the salary levels for employees who quit.
3. Create a combined bar chart for employee turnover by department and salary.

Specify the Question

1. Which department had the highest number of employees quit?
2. What was the breakdown of employees quitting by salary level?

Obtain the Data

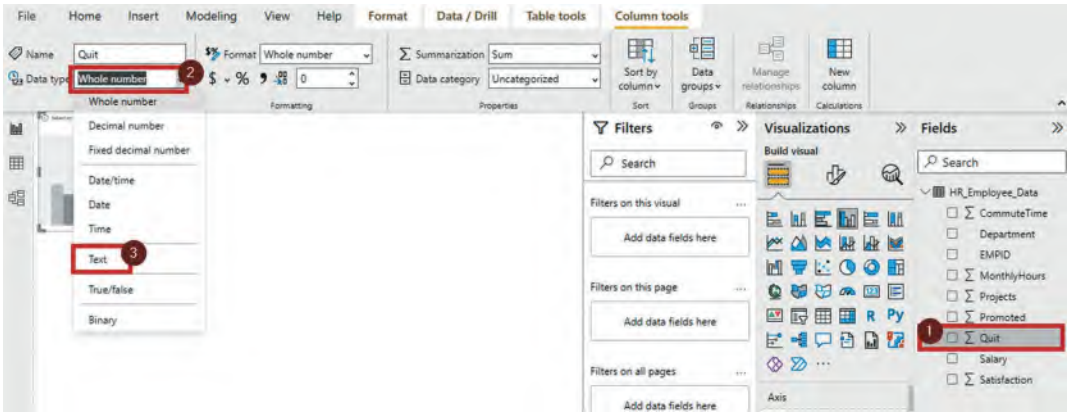
This lab uses the data in Lab 10.2 Data.xlsx. The human resource (HR) data are from Kaggle.com and based on a real multinational company that we will call MNC Inc. The data dictionary for this file is below.

Data Dictionary

EMPID:	Unique employee ID
Satisfaction:	Satisfaction level of the employee (0–1: 0 = not satisfied, 1 = very satisfied)
Projects:	Number of projects that an employee is currently assigned to
MonthlyHours:	Average hours worked by the employee in the last three months
CommuteTime:	Monthly commute time to office
Quit:	1 = employee quit, 0 = employee is still employed by MNC
Promoted:	1 = employee promoted within the past five years, 0 = no promotion in past five years
Department:	Department employee works in
Salary:	Relative salary paid to employee (low, medium, high)

Obtain Task 1: Load the Data into Power BI.

1. Open Power BI.
2. Click Get Data, then verify Excel Workbook is highlighted in the Get Data dialog box, and click Connect.
3. In the Open window, browse to the Lab 10.2 Data.xlsx file.
4. Click Open.
5. Place a check mark next to HR_Employee_Data and click Load.
6. Power BI defined Quit as a number (see the sigma symbol next to the field in the expanded HR_Employee_Data table in the Fields pane). Because Quit indicates a category of employees (Quit, Not Quit), not a number used in computations, change Quit to a string. To do this, select the Quit field from the Fields pane and adjust the Data Type from Whole number to Text, as shown in Exhibit 10.2P.1.



Lab Exhibit 10.2P.1
Microsoft Power BI

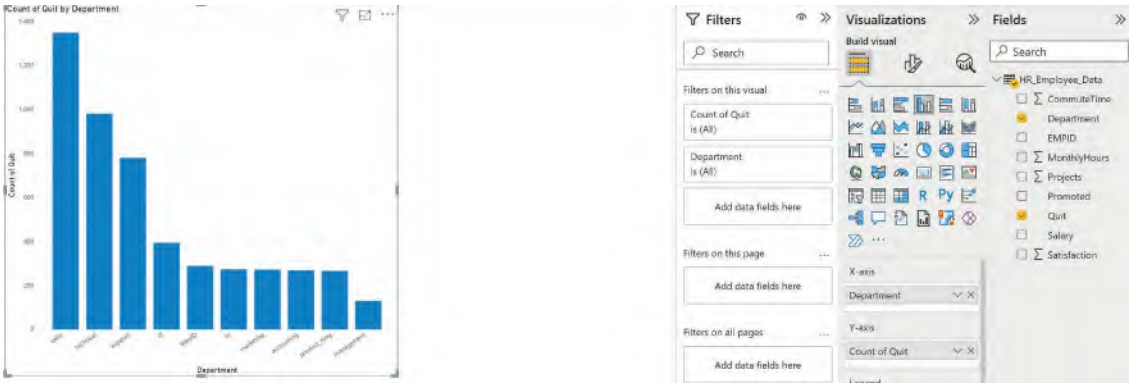
7. Repeat the steps to set Promoted to string data. The data are now ready to use.

Analyze the Data

Analysis Task 1: Create the First Bar Chart: Number of Employees Leaving by Department.

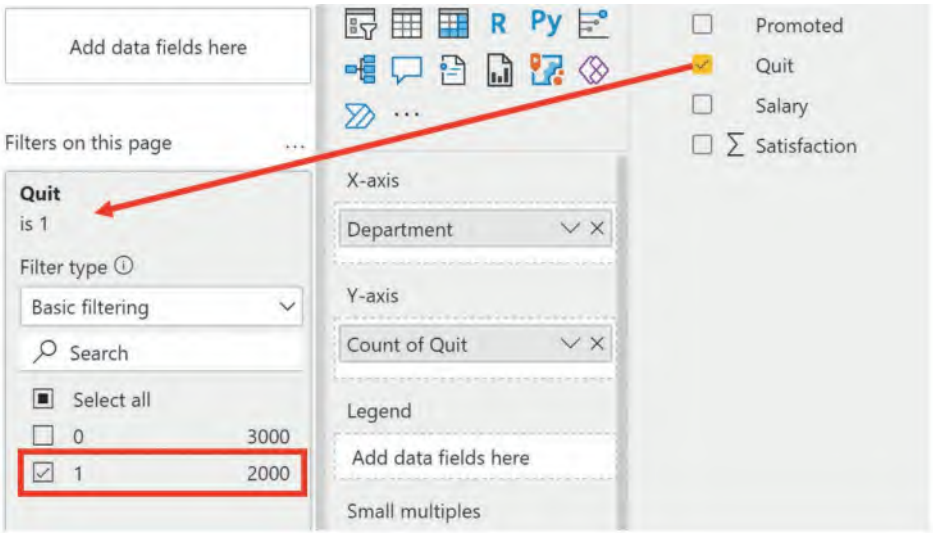
In the first bar chart, we will determine the number of employees who quit in each department.

1. Click Clustered column chart from the Visualizations pane in the Build visual tab.
2. Drag Department to the box under X-axis that says Add data fields here in the Visualizations pane (with the Build visual tab selected).
3. Drag Quit to the box under Y-axis that says Add data fields here in the Visualizations pane (with the Build visual tab selected). Your visualization should look like Lab Exhibit 10.2P.2.



Lab Exhibit 10.2P.2
Microsoft Power BI

4. Currently, the visualization shows all the values of Quit, so it counts both 0s and 1s. We want only the employees leaving the company to show, or Quit = 1. We need to create a Filter to do this. To do so, drag Quit from the Fields pane to the Filters on this page shelf in the Filters pane where it says Add data fields here. The Filter Field [Quit] dialog box appears. Select 1, as shown in Lab Exhibit 10.2P.3.



Lab Exhibit 10.2P.3

Microsoft Power BI

5. To make the bar chart more precise, let's label each column with the exact number of employees quitting. In the Visualizations pane, select the Format (your) visual icon and toggle the Data labels from Off to On.
6. Let's also rename the Y-axis. In the Format visual tab of the Visualizations pane, expand the Y-axis category, and make sure the Title is On. Expand the Title. Replace the text "Auto" in the Title text box with "Number of Employees who Quit". Unexpand the Y axis category.
7. Before moving to the next bar chart, let's name the data visualization. In the Format (your) visual icon, General tab of the Visualization pane, expand Title, and replace the auto Text with "Number of Employees who Quit by Department."

Analysis Task 2: Create the Second Bar Chart: Salary Level of Employees Quitting.

In the second bar chart, we will break down the number of employees who quit by salary level (low, medium, high).

1. Click on the + at the bottom of the visualization to open a new Page.
2. Click Clustered column chart from the Visualizations pane (with the Build visual icon selected).
3. Drag Salary from the Fields pane to the Add data fields here box under X-axis in the Visualizations pane (with the Build visual icon selected).
4. Drag Quit from the Fields pane to the Add data fields here box under Y-axis in the Visualizations pane (with the Build visual icon selected).
5. Set a filter to display just employees who quit following the instructions Analysis Task 2.
6. Label the columns with the precise number following the procedure outlined in Analysis Task 2.
7. Change the Y-axis label to "Number of Employees who Quit".
8. Name the data visualization "Number of Employees who Quit by Salary Level".

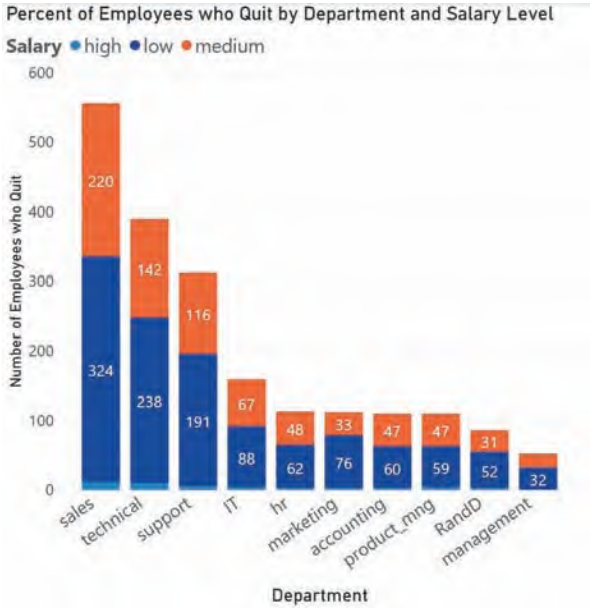
Report the Results: Create a Combined Bar Chart

You have just calculated (1) the number of employees who quit in each department and (2) the number of employees who quit in each salary level. Management was not very specific about how it wanted the results reported. So, to help report the results to management, let's create a combined visualization of this information using percentages. Follow these steps:

1. Create a copy of the Numbers of Employees who Quit by Department tab. (Right-click on the chart, select **Copy > Copy visual**, and paste it using Ctrl-V on a new page.)
2. Double check that the Filter for employees who quit is still active. If the filter is missing, add it to the visual.
3. Change the title to "Percent of Employees who Quit by Department and Salary Level".
4. Drag Salary from the Fields pane to the box that says Add data fields here under Legend in the Visualizations pane with the Build visual icon selected.
5. Adjust the visualization type to a Stacked column chart.
6. If necessary, add data labels by using the Format (your) visual icon (Visual tab) of the Visualizations pane and toggling Data labels to the On position. Your data visualization should look like Lab Exhibit 10.2P.4.

Lab Exhibit 10.2P.4

Microsoft Power BI



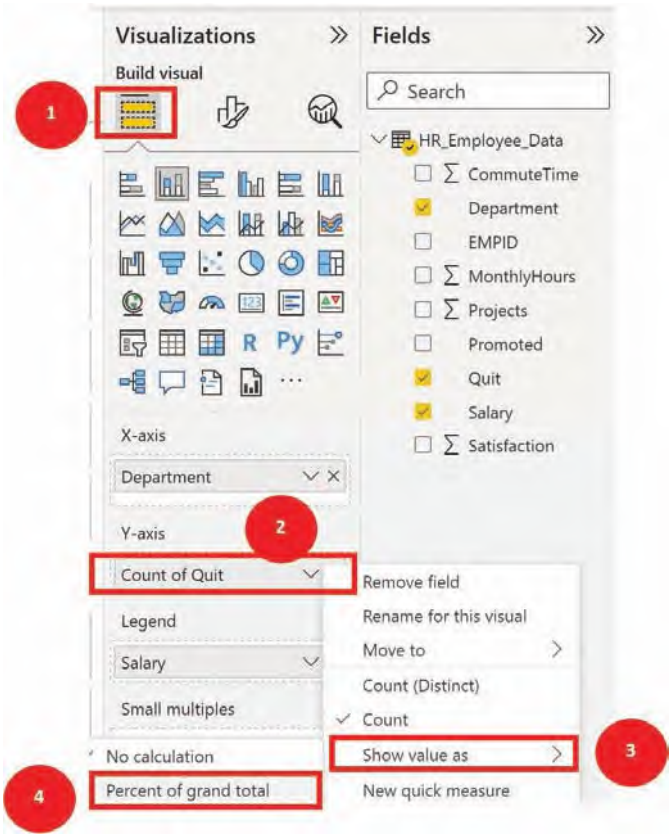
7. To see the percentage of the total number of employees who quit, proceed as follows. In the Build visual option of the Visualizations pane, click the down arrow by Count of Quit field in Y-axis, and select **Show value as > Percent of grand total** as shown in Lab Exhibit 10.2P.5.
8. To remove the decimals from the percentages, proceed as follows: In the Visualizations pane Format (your) visual option (Visual tab), expand the Data labels option. Expand the Values option. Set Value decimal places to 0.

Assessment

1. Take a screenshot of the bar chart showing the number of employees who quit by department, copy it into a Word document named "Lab 10.2 Power BI Submission.docx", and label the screenshot Submission 1. Take a screenshot of the completed stacked bar chart for the number of employees who quit by salary level, paste it into the same Word document, and label the screenshot Submission 2. Take a screenshot

of the completed stacked bar chart for the number of employees who quit by department and salary level, paste it into the same Word document, and label the screenshot Submission 3.

2. Answer the questions in Connect and upload your Lab 10.2 Power BI Submission.docx via Connect, if assigned.



Lab Exhibit 10.2P.5

Microsoft Power BI

Alternate Lab 10.2

Descriptive Analytics/Human Resources: Analyzing Employee Turnover Using a Bar Chart

Using either Tableau or Power BI, apply the same steps in Lab 10.2 to the Alt Lab 10.2 Data.xlsx data set. For decision-making context, please see the start of Lab 10.2.

Required

1. Create a bar chart breaking down employee turnover by department.
2. Create a bar chart showing the salary levels for employees who quit.
3. Create a combined bar chart for employee turnover by department and salary.

Specify the Question

1. Which department had the highest number of employees quit?
2. What was the breakdown of employees quitting by salary level?

Assessment

1. Take a screenshot of the bar chart showing the number of employees who quit by department, copy it into a Word document named "Alt Lab 10.2 Tableau [or Power BI] Submission.docx", and label the screenshot Submission 1. Take a screenshot of the completed stacked bar chart for the number of employees who quit by salary level, paste it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the completed stacked bar chart for the number of employees who quit by department and salary level, paste it into the same Word document, and label the screenshot Submission 3.
2. Answer the questions in Connect and upload your Alt Lab 10.2 Tableau/Power BI Submission.docx via Connect, if assigned.

Lab 10.3 Excel

Diagnostic Analytics/Supply Chain: Analyzing KPI Differences Between Walmart and Amazon

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Diagnostic Analytics, KPIs, Hypothesis Test, Supply Chain

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Diagnostic analytics attempt to find the root cause by answering the questions "Why did it happen?" and "What are the causes of past results?" Hypothesis testing is a diagnostic analytics technique used to test whether variables are significantly different. In this lab, you will test whether selected supply chain KPIs are different for **Amazon** and **Walmart** during the period 1999–2021. Managers may perform this type of analysis to see how their company's operations compare to other companies in the industry and to identify areas for improvement.

Required

1. Perform hypothesis tests to determine if selected supply chain KPIs (asset turnover, inventory-to-sales ratio, inventory turnover) are different between Amazon and Walmart.
2. Create a scatterplot for asset turnover values from 1999 to 2021 for both Amazon and Walmart.

Specify the Question

Are selected supply chain KPIs (asset turnover, inventory-to-sales ratio, inventory turnover) for Amazon and Walmart different?

Obtain the Data

This lab uses the KPIs from Lab 10.1, which you can find in the Lab 10.3 Data.xlsx file. The data dictionary for this file is below. The data include annual financial information for the years 1999 to 2021.[29] Each KPI has its own (labeled) worksheet.

Data Dictionary

Year: Data (fiscal) year
KPI for Amazon: KPI (asset turnover, inventory-to-sales, inventory turnover)
KPI for Walmart: KPI (asset turnover, inventory-to-sales, inventory turnover)

Analyze the Data

Analysis Task 1: Install the Analysis ToolPak.

1. This analysis uses the Analysis ToolPak for Excel. If you have not installed the ToolPak, please see textbook Appendix E for installation instructions.

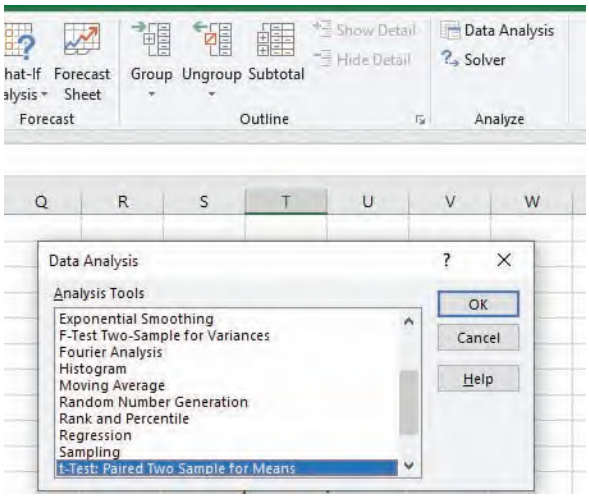
[29]Wharton Research Data Services (WRDS), "Compustat North American Fundamentals—Annual." https://wrds-www.wharton.upenn.edu/ 3/13/2022.

Analysis Task 2: Generate a *t*-test Using the Analysis ToolPak.

Notice that the data for Amazon and Walmart have been organized in a special way for hypothesis testing in Excel. You will use a paired two sample for means because the KPI for each company was measured at the same time (each year). Each column contains data for one company, and each row contains data measured in a particular year. You will start with the Asset Turnover worksheet.

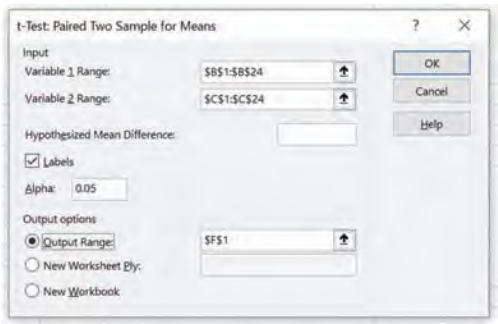
1. Select **Data > Analyze > Data Analysis.**
2. Select t-test: Paired Two Sample for Means, as shown in Lab Exhibit 10.3.1.
3. Click OK.

Lab Exhibit 10.3.1



4. The t-Test: Paired Two Sample for Means dialog box should appear. In the Variable 1 Range box, type "$B$1: $B$24". In the Variable 2 Range box, type "$C$1:$C$24". Check the Labels box because the labels are included in the data (that is, B1 and C1). Accept the Alpha as 0.05, which means that we will be 95 percent confident in the test results. Because there are multiple KPI worksheets, put the results on the Asset Turnover worksheet. To do this, select Output Range and type $F$1 in the associated box. Lab Exhibit 10.3.2 shows what the t-Test: Paired Two Sample for Means dialog box should look like. Click OK.

Lab Exhibit 10.3.2



Analysis Task 3: Interpret the Hypothesis Test Results.

Lab Exhibit 10.3.3 presents the results of the hypothesis test. Over the period 1999–2021, the average (**Mean**) asset turnover ratios were 1.94 and 2.45 for Amazon and Walmart, respectively. The t Stat for the data is −4.26.

Lab Exhibit 10.3.3

t-Test: Paired Two Sample for Means		
	AMAZON.COM INC	*WALMART INC*
Mean	1.944443908	2.451195445
Variance	0.28616868	0.019315946
Observations	23	23
Pearson Correlation	-0.135432847	
Hypothesized Mean Difference	0	
df	22	
t Stat	-4.258939758	
P(T<=t) one-tail	0.000160275	
t Critical one-tail	1.717144374	
P(T<=t) two-tail	0.000320551	
t Critical two-tail	2.073873068	

Because we did not make a directional prediction (that is, we did not predict whether Amazon's asset turnover ratio is higher or lower than Walmart's asset turnover ratio), the *p*-value in which we are interested is the P two tail, which equals 0.0003. A small *p*-value indicates that the difference is significant. In addition, Excel also shows the t Critical two-tail value as 2.07 (which is based on the alpha of 0.05 that we set in the t-Test: Paired Two Sample for Means dialog box). If the absolute value of the t Stat (4.26) is greater than the critical value (2.07), then the difference is significant. These results are highlighted in yellow in Lab Exhibit 10.3.3.

If we made a directional prediction, we could use the P one-tail and the t Critical one-tail tests. If we hypothesized that Walmart had the higher asset turnover, then the t Stat would still be −4.26, which is still larger than the t Critical one-tail value (1.72), indicating that Walmart's average is significantly higher than Amazon's. Note the p-value is small at 0.0002.

Based on our hypothesis test, Walmart has averaged a higher asset turnover ratio for the 1999–2021 period, indicating that it is more efficient at generating revenues from its assets.

Analysis Task 4: Perform Hypothesis Test on the Remaining KPIs.

Use the steps above to perform hypothesis tests on the remaining KPIs (inventory-to-sales ratio and inventory turnover).

Report the Results: Plot Asset Turnover Over the Years


DATA VISUALIZATION

You have just run hypothesis tests to see if select KPIs for Amazon and Walmart are different during the 1999–2021 period. An important role for diagnostic analytics is helping decision-makers understand past results. Let's plot the annual values for asset turnover for both companies to see how the ratios change over time.

1. In the Asset Turnover tab, highlight the data (A1:C24).
2. From the menu, select **Insert > Charts > Insert Scatter (X, Y) or Bubble Chart > Scatter.**
3. Rename the chart "Scatter Plot of Asset Turnover".
4. From the menu, select **Chart Tools > Design > Add Chart Element > Axis Titles > Primary Vertical.**
5. Change the Y-axis Axis Title to "Asset Turnover".
6. Change the X-axis Axis Title to "Year".

Lab Exhibit 10.3.4 shows the scatterplot. This data visualization shows how asset turnover has changed over the years for both companies. The plot clearly shows that Walmart's asset turnover ratio has had a slight decline over the years, but it was fairly stable over time. In contrast, Amazon's asset turnover ratio changes over time. This data visualization provides great context for the hypothesis tests.

Lab Exhibit 10.3.4



Assessment

1. Take a screenshot of the *t*-test for asset turnover, copy it into a Word document named "Lab 10.3 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the *t*-test for inventory-to-sales, paste it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the *t*-test for inventory turnover, paste it into the same Word document, and label the screenshot Submission 3. Take a screenshot of the asset turnover scatterplot, paste it into the same Word document, and label the screenshot Submission 4.
2. Answer the questions in Connect and upload your Lab 10.3 Excel Submission.docx file via Connect, if assigned.

Alternate Lab 10.3 Excel

Diagnostic Analytics/Supply Chain: Analyzing KPI Differences Between Bed, Bath, and Beyond and Target

Apply the same steps in Lab 10.3 to the Alt Lab 10.3 Data.xlsx data set.

Required

1. Perform hypothesis tests to see if selected supply chain KPIs (asset turnover, inventory-to-sales ratio, inventory turnover) are different between **Bed, Bath, and Beyond** and **Target.**
2. Create a scatterplot for asset turnover values from 1999 to 2019 for both Bed, Bath, and Beyond and Target.

Specify the Question

Are selected supply chain KPIs (asset turnover, inventory-to-sales ratio, inventory turnover) for Bed, Bath, and Beyond and Target different?

Assessment

1. Take a screenshot of the *t*-test for asset turnover, copy it into a Word document named "Alt Lab 10.3 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the *t*-test for inventory-to-sales, paste it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the *t*-test for inventory turnover, paste it into the same Word document, and label the screenshot Submission 3. Take a screenshot of the asset turnover scatterplot, paste it into the same Word document, and label the screenshot Submission 4.
2. Answer the questions in Connect and upload your Alt Lab 10.3 Excel submission.docx file via Connect, if assigned.

Lab 10.4 Tableau

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Diagnostic Analytics/IT Operations: Analyzing Website Response Time Using a Control Chart

Keywords

Diagnostic Analytics, Tableau, Control Chart, IT Operations

Decision-Making Context

Understanding what happened in the past is an important first step in harnessing the power of business analytics. Diagnostic analytics attempt to find the root cause by answering the questions "Why did it happen?" and "What are the causes of past results?" As described in the chapter, control charts are one technique used in diagnostic analytics to determine if data collected over time contain statistically significant signals or if the variation is just noise.

In this lab, you will create a control chart for slow website response time using Tableau. FastnFurious Company has the goal of a three-second response time because users will leave its website if the website is not responsive. Unfortunately, user engagement time on the website has decreased, causing profits to drop. Based on customer complaints, managers know that average website response time has increased. Managers are trying to determine exactly when the response time increased so they can figure out the root cause. Managers want to start by focusing on response time greater than three standard deviations from the mean. To start their analysis, they want a weekly control chart developed for the past 52 weeks.

Required

Create a control chart for website response time data over the past 52 weeks.

Specify the Question

When did average website response time begin to increase beyond acceptable times?

Obtain the Data

This lab uses the data in Lab 10.4 Data.xlsx. The file consists of 52 weeks of slow response time counts for a company website. The data dictionary for this file is below.

Data Dictionary

Slow Response Time:	Number of times during the week that the company website took more than three seconds to respond
Week:	Week number

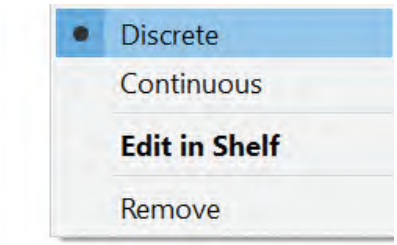
Obtain Task 1: Load the Data into Tableau.

1. Open Tableau.
2. Under Connect, select Microsoft Excel.
3. In the Open window, browse to the Lab 10.4 Data.xlsx file.
4. Click Open. The data are now ready to use.

Analyze the Data

Analysis Task 1: Create a Process Control Chart—Add Lines.

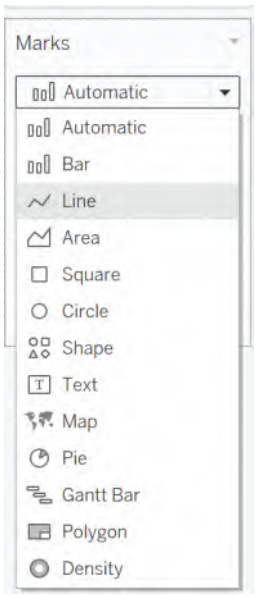
1. Click on Sheet 1.
2. Drag Week from the Data pane to the Columns shelf.
3. Click the down arrow in the Week pill and select Discrete to change the type from Continuous to Discrete, as shown in Lab Exhibit 10.4.1. The week pill should turn blue (which indicates a discrete variable). Discrete variables are whole numbers that can be put into bins or categories, while continuous variables can take on any (fractional) value. Because we want to analyze the data by week, discrete is the proper choice to divide the data into categories for each week.



Lab Exhibit 10.4.1

Tableau Software

4. Drag Slow Response Time from the Data pane to the Rows shelf.
5. If a bar chart appears, select Line in the Marks card to change it to a line chart, as shown in Lab Exhibit 10.4.2.



Lab Exhibit 10.4.2

Tableau Software

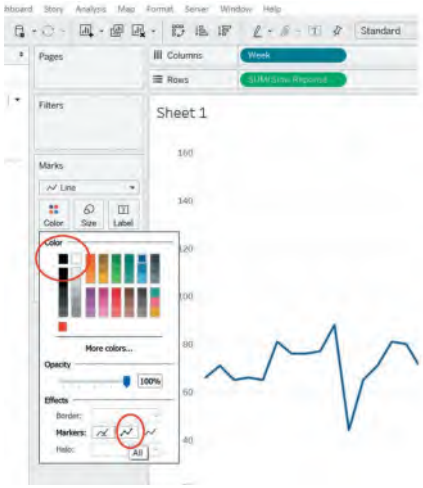
Analysis Task 2: Create a Control Chart—Add Data Points.

Let's make the line black and add circles for the data points.

1. Click Color in the Marks card.
2. To make the line black, click on the color at the top-left corner.
3. In the Color dialog box that appears, click the middle Markers line, which may be labeled "All," as shown in Lab Exhibit 10.4.3. The line should now have circles at each data point.

Lab Exhibit 10.4.3

Tableau Software



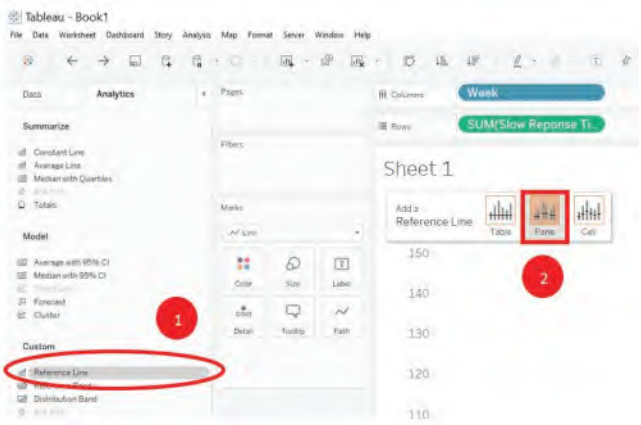
Analysis Task 3: Create a Control Chart—Add the Average Line.

We are now going to add the average line, upper control limit line, and the lower control limit line. These steps create a blue dashed line for the Average.

1. Click on the Analytics tab (next to the Data tab at the top of the Data pane).
2. Drag Reference Line to the View (where the visualization is) and then drag it over the Pane picture in the Add a Reference Line dialog box, as shown in Lab Exhibit 10.4.4. If you are given a choice of variables, select SUM(Slow Response Time).
3. The Edit Reference Line, Band, or Box dialog box should open. Ensure that the Line box at the top is highlighted and the Line Value is set to SUM(Slow Response Time).

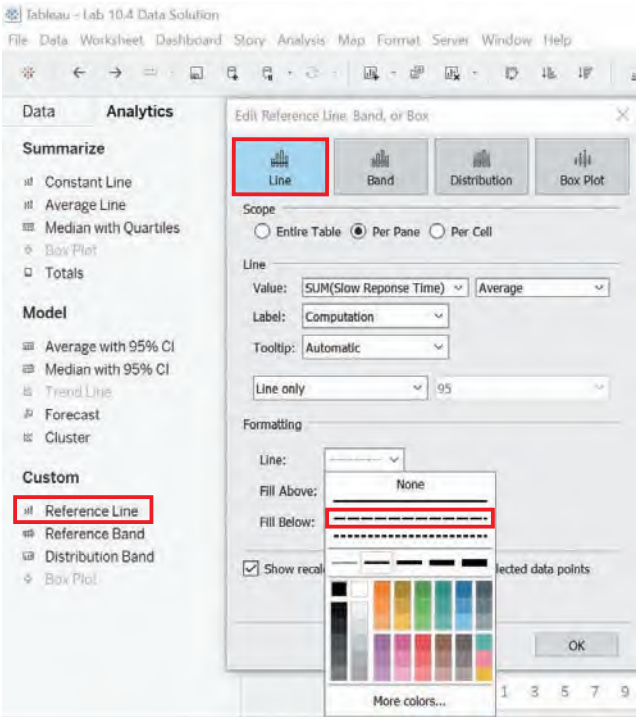
Lab Exhibit 10.4.4

Tableau Software



Accept the remaining Line section values. In the Formatting section, click on Line and select the first dashed line.

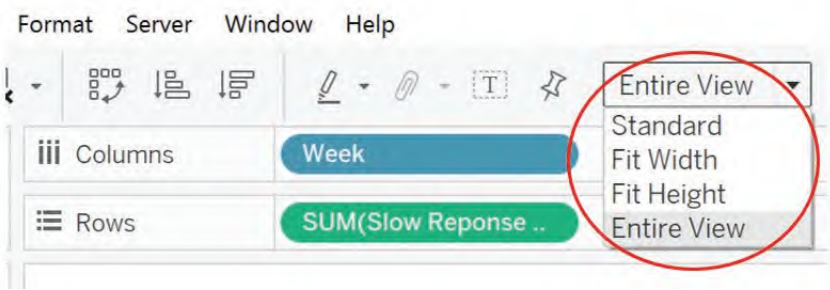
4. Select blue for the line color, as shown in Lab Exhibit 10.4.5.
5. Click OK.



Lab Exhibit 10.4.5

Tableau Software

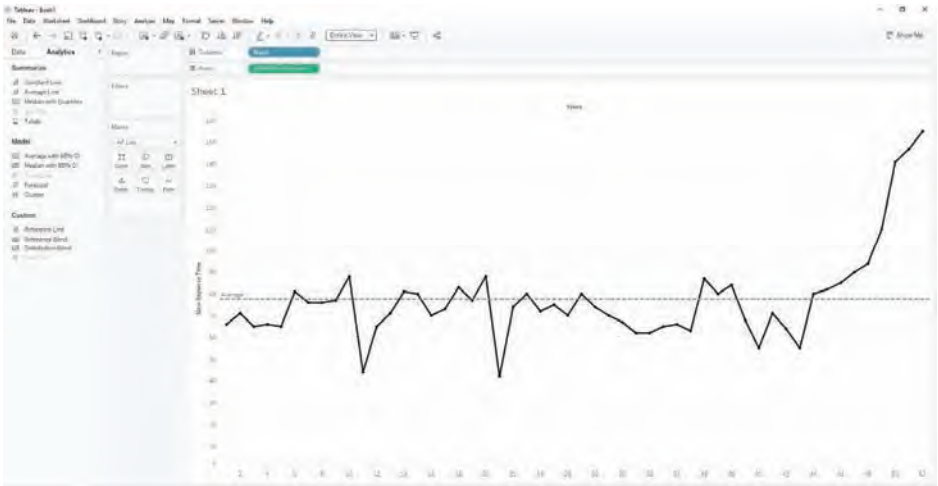
6. If you cannot see the entire data set in the View without having to scroll or if the line does not completely fill the View, change the view from Standard to Entire View, as shown in Lab Exhibit 10.4.6.



Lab Exhibit 10.4.6

Tableau Software

The visualization should look like Lab Exhibit 10.4.7.



Lab Exhibit 10.4.7

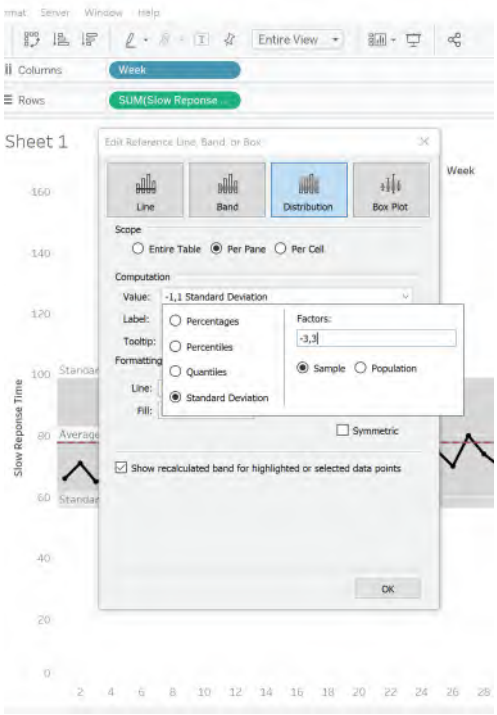
Tableau Software

Analysis Task 4: Create a Control Chart—Add the Upper and Lower Control Limit Lines.

1. Click on the Analytics tab (next to the Data tab at the top of the Data pane).
2. Drag Reference Band to the View (where the lines are) and then drag it over the Pane picture in the Add a Reference Band dialog box (similar to what we did in Analysis Task 4). If you are given a choice of variables, select SUM(Slow Response Time).
3. In the Edit Reference Line, Band, or Box dialog box, click on Distribution.
4. In the Computation section, click on the down arrow by Value.
5. Select Standard Deviation.
6. In the Factors: box, type "−3, 3" to indicate that three standard deviations are to be used for the upper and lower control limit lines, the UCL and LCL, respectively. Values outside these control limit lines are outliers, which should be investigated. Lab Exhibit 10.4.8 shows what the Edit Reference Line, Band, or Box should look like.

Lab Exhibit 10.4.8

Tableau Software



7. Click the down arrow again to close the Value box.
8. In the Formatting section, click the down arrow by Line.
9. Select the first dashed line.
10. Change the line color to red, which will make the UCL and LCL dashed red lines.
11. Click OK.
12. Double-click on the Sheet 1 tab, type "Slow Response", and then press Enter. The control chart should look like Lab Exhibit 10.4.9.



Lab Exhibit 10.4.9

Tableau Software

Analysis Task 5: Interpret a Control Chart

The dashed blue line represents the average number of response times that exceeded three seconds during a week. Because managers of FastnFurious Company want response time to be as fast as possible, their focus right now is on the UCL, represented by the dashed upper red line at 140. Beginning in week 50, the number of response times longer than three seconds exceeded three standard deviations above the Average dashed blue line. Managers need to determine what caused the increase beginning in week 50. In the long run, management would like all response times to be less than three seconds. Management could pursue a Six Sigma strategy to gradually optimize performance, slowly moving down the Average, the UCL, and the LCL (first by 2.5 standard deviations, then by 2 standard deviations, then by 1.5 standard deviations, and so on). The ultimate goal for FastnFurious is to lower the Average as close as possible to zero with a very narrow LCL and UCL.

Report the Results: Create Six Sigma Control Charts

We just created a control chart for website slow response times over the last 52 weeks. After looking at the chart, FastnFurious management would like two additional control charts as it attempts to implement Six Sigma in the company: one chart that uses two standard deviations for the upper and lower control limit lines and a second chart that uses one standard deviation for the upper and lower control limit lines. Create these two additional control charts on separate Tableau tabs using the following steps.

1. Click on the Analytics tab.
2. Drag Reference Band to the View and then drag it over the Pane in the Add a Reference Band dialog box. If you are given a choice of variables, select SUM(Response Time).
3. In the Edit Reference Line, Band, or Box dialog box, click on Distribution.
4. In the Computation section, click on the down arrow by Value.
5. Select Standard Deviation.
6. In the Factors box, type "−2, 2".
7. Click the down arrow again to close the Value box.
8. In the Formatting section, click the down arrow by Line.
9. Select the first dashed line.
10. Change the line color to orange, which will make the 2-standard-deviation UCL and LCL dashed orange lines.

Six Sigma: A set of techniques and tools designed to achieve continual process improvement almost to the point of zero defects. Six Sigma is a reference to the three standard deviations on either side of the mean of a normally distributed variable.

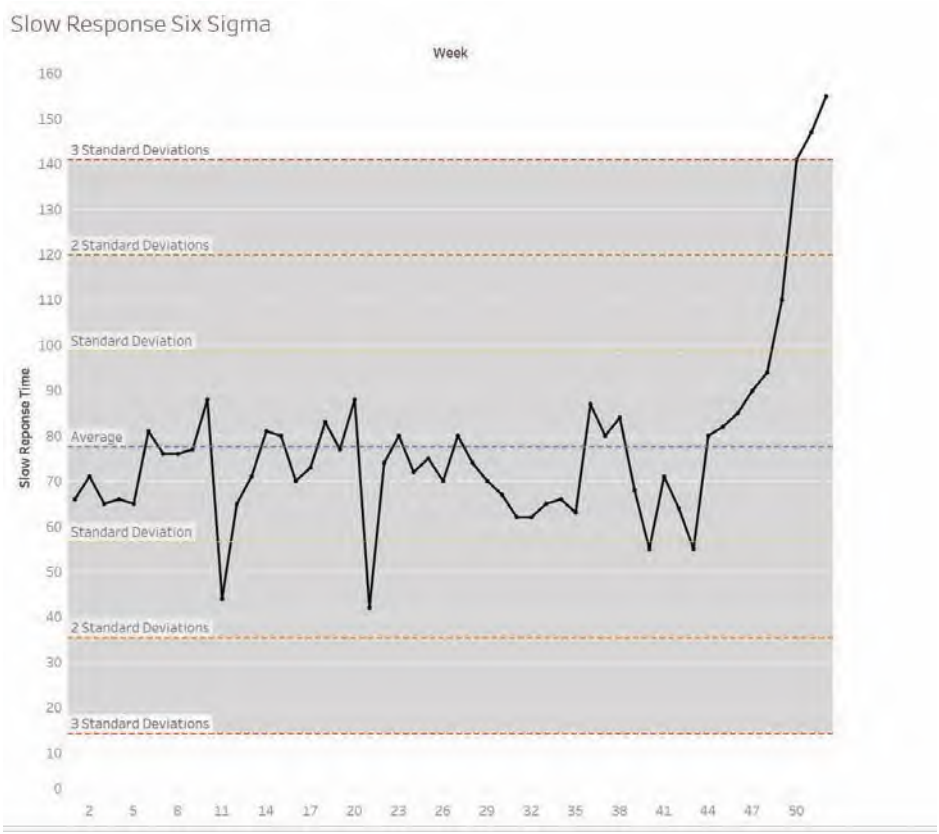


DATA VISUALIZATION

11. Click OK.
12. Repeat steps 1–11 using "−1,1" as the standard deviations and yellow as the color.

Lab Exhibit 10.4.10

Tableau Software



13. Change the Y-axis label to Number of Slow Response Times. To do so, double-click on Y-axis label. Change the title to "Number of Slow Response Times". Click x at the top-right corner to close the Edit Axis dialog box.

Lab Exhibit 10.4.10 shows what the data visualization should look like. How well does the company perform with smaller control limits?

Assessment

1. Take a screenshot of the control chart, and copy it into a Word document named "Lab 10.4 Tableau Submission.docx".
2. Answer the questions in Connect, and upload your Lab 10.4 Tableau Submission.docx file via Connect, if assigned.

Alternate Lab 10.4 Tableau

Diagnostic Analytics/IT Operations: Analyzing Website Response Time Using a Control Chart

Apply the same steps in Lab 10.4 to the Alt Lab 10.4 Data.xlsx data set. For decision-making context, please see the start of Lab 10.4.

Required

Create control charts for website response time data over the past 52 weeks.

Specify the Question

When did average website response time begin to increase beyond acceptable times?

Assessment

1. Take a screenshot of the control chart, and copy it into a Word document named "Alt Lab 10.4 Tableau Submission.docx".
2. Answer the questions in Connect and upload your Alt Lab 10.4 Tableau Submission.docx file via Connect, if assigned.

Lab 10.5 Tableau

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Predictive Analytics/Supply Chain: Forecasting Unit Demand[30]

Keywords

Predictive Analytics, Time Series, Trend Line, Forecast, Supply Chain

Decision-Making Context

As described in this chapter, predictive analytics attempts to answer the questions "What will happen in the future?" and "What is the probability something will happen?" and "Can we forecast what will happen?" The goal is to predict future performance using historical data. So, unlike descriptive and diagnostic analytics, which describe things we know are true, predictive analytics describes things that possibly might happen.

This lab looks at two types of predictive analytics techniques, regression and forecast, using a time series. A *time series* is a series of data points ordered in time. For example, the quarterly or annual reporting of a company's financial information is an example of a time series. Time series can be used to create regressions and forecasts. A *regression* predicts a specific outcome based on the relationship between independent and dependent (outcome) variables. A *forecast* predicts future values based on past values of the same variable. Forecasts anticipate the behavior of a variable on long timelines. In this lab, you will create a regression for a time series of unit demand as well as a unit demand forecast for Don't Fall Inc., a fictional hoverboard company. Don't Fall Inc. sells only one type of hoverboard, but it has the best features and durability of all hoverboards on the market.

Required

Create a regression (call a Trend Line in Tableau) and create a forecast for hoverboard unit demand.

Specify the Question

What is the relationship between time and unit demand? What is expected unit demand for the next year?

Obtain the Data

This lab uses the data in Lab 10.5 Data.xlsx. The file consists of 49 quarters of (unit) product demand starting in the first quarter of 2021 and ending in the fourth quarter of 2024. The data dictionary for this file is below.

Data Dictionary

Quarter: The quarter product demand was reported (YYYYQ#, #=1–4)
Demand: Demand reported for the quarter

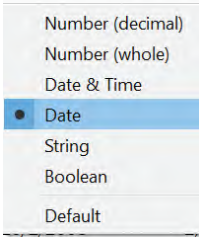
[30]This lab is a very simplified version of an article by Farzan Sajahan, January 18, 2021, "Time Series Analysis in Tableau." Available online at https://dataschoolonline.com/time-series-analysis-in-tableau/. Accessed 7/22/2021. Please see this link for more in-depth information about time series trend lines and forecasts.

Obtain Task 1: Load the Data into Tableau.

1. Open Tableau.
2. Under Connect, select Microsoft Excel.
3. In the Open window, browse to the Lab 10.5 Data.xlsx file. Click Open.

Obtain Task 2: Prepare the Data.

1. Tableau reads Quarter as a string (that is, text). Tableau requires at least one date dimension and one measure to create a forecast. So, change Quarter to a Date. To do this, click on the Abc above Quarter.
2. Select Date, as shown in Lab Exhibit 10.5.1. The data are now ready to use.



Lab Exhibit 10.5.1

Tableau Software

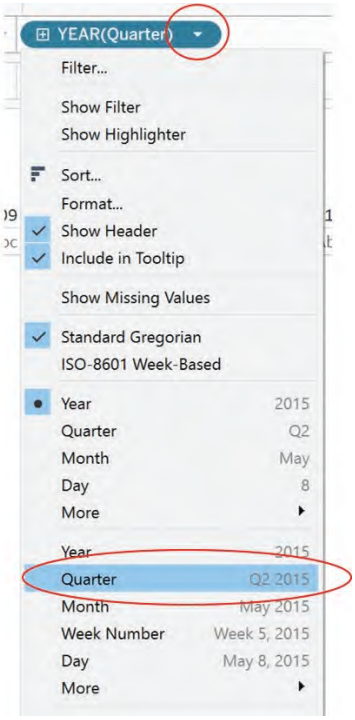
Analyze the Data

Analysis Task 1: Graph the Time Series.



DATA VISUALIZATION

1. Click on Sheet 1.
2. Drag Quarter from the Data pane to the Columns shelf. If the Quarter pill says YEAR, click on the arrow in the Quarter pill and select the second Quarter (Q2 2015) to change to QUARTER, as shown in Lab Exhibit 10.5.2.



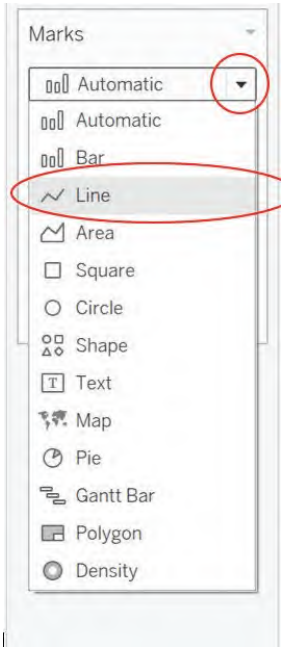
Lab Exhibit 10.5.2

Tableau Software

3. Drag Demand from the Data pane to the Rows shelf. If a bar chart appears, in the Marks card use the down arrow to change Automatic to Line, as shown in Lab Exhibit 10.5.3.

Lab Exhibit 10.5.3

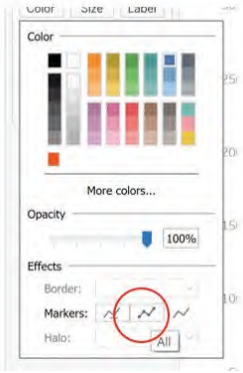
Tableau Software



4. Let's add circles for each data point. Click Color in the Marks card.
5. In the Effects section in the Markers line, click the second marker (which is labeled All), as shown in Lab Exhibit 10.5.4.

Lab Exhibit 10.5.4

Tableau Software

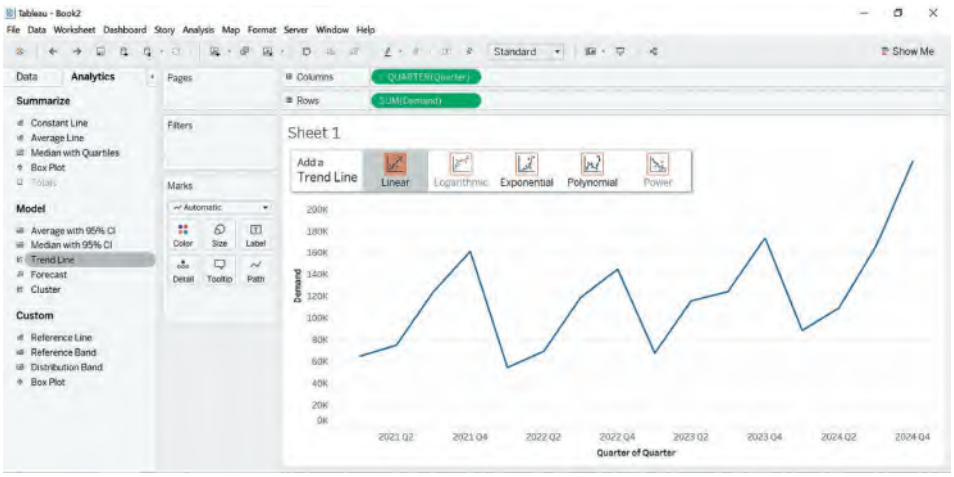


6. Click anywhere outside of the Color dialog box to close it.
7. Before moving on, let's look at the graph. The data points make it easy to see that the company experiences seasonality. *Seasonality* is a characteristic of time series where the data exhibit predictable changes that occur at specific regular intervals. It appears that demand drops in the first quarter and peaks in the fourth quarter.

Analysis Task 2: Add a Trend Line.

Now, let's run a regression, which Tableau executes via the *trend line.*

1. Click on the Analytics tab (next to the Data tab at the top of the Data pane).
2. Drag Trend Line to the View (where the visualization is) and then drag it over the Linear picture in the Add a Trend Line dialog box, as shown in Lab Exhibit 10.5.5. The trend line is based on the product demand data points in the time series.

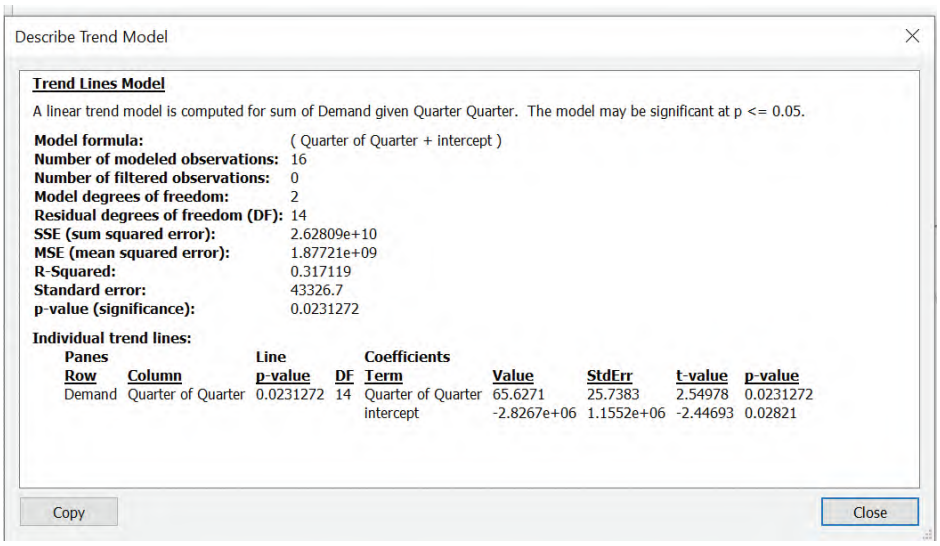


Lab Exhibit 10.5.5

Tableau Software

Analysis Task 3: Interpret the Trend Line.

Putting the cursor on the line reveals the trend line equation, R-squared, and p-value. For more information, right-click on the Trend Line, select Describe Trend Model (or **Analysis > Trend Lines > Describe Trend Model**). The Describe Trend Model dialog box opens and looks like Lab Exhibit 10.5.6.



Lab Exhibit 10.5.6

Tableau Software

Examine the output in the Describe Trend Model dialog box. Because this is a forecast, let's skip to the Individual trend lines section. The coefficients that appear at the bottom of the Describe Trend Model dialog box provide information about Quarter (of Quarter) and the intercept. The t-value for Quarter (of Quarter) (2.54978) has a p-value (0.0231272) less than 0.05, which indicates that there is a significant trend in the data, or a positive relationship between Quarter (of Quarter) and Demand. That is, Demand increases with time. The intercept is also significant.

The coefficients are the regression coefficients and are used to determine the regression equation (or fitted line plot). The regression equation is:

$$y = 65.63x + (-2.83e{+}6)$$

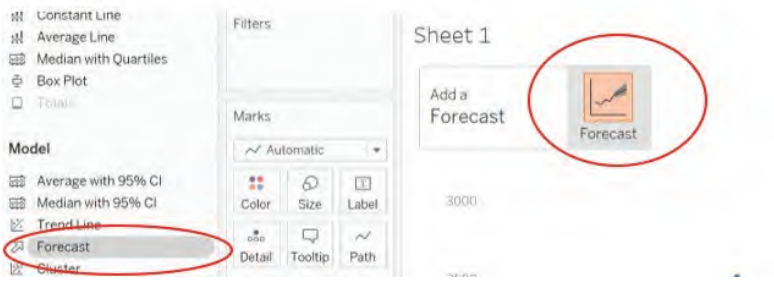
This regression equation indicates that demand increases on average by 65.63 each quarter after starting near zero (−2.83 e+6) on the Y-axis intercept. So, the trend line has just smoothed out the seasonality in the data into an upward trend.

Analysis Task 4: Add the Forecast.

1. Click on the Analytics tab (next to the Data tab at the top of the Data pane).
2. Drag Forecast to the View (where the visualization is) and then drag it over the Forecast picture in the Add a Forecast dialog box, as shown in Lab Exhibit 10.5.7.

Lab Exhibit 10.5.7

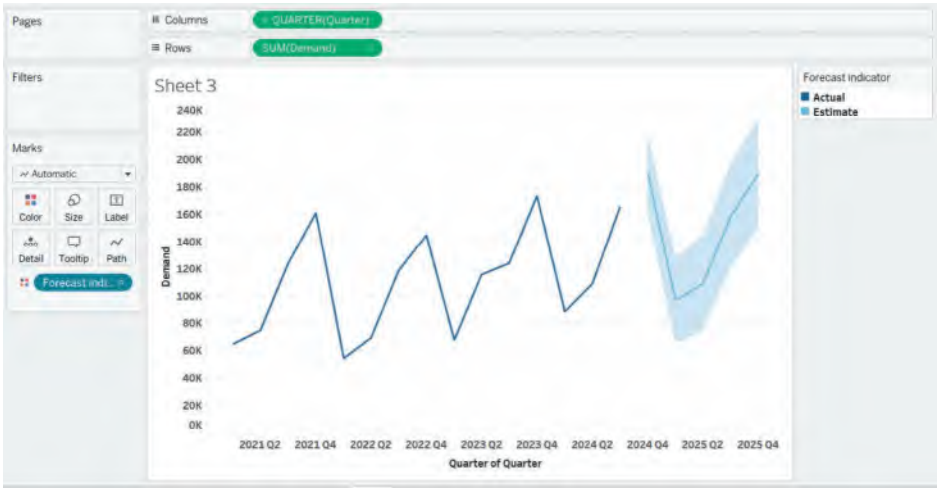
Tableau Software



3. Tableau automatically created a one-year forecast through the fourth quarter of 2025, which is shaded to differentiate it from the actual data. Tableau also created a new trend line for the forecasted data (Lab Exhibit 10.5.8). If you hover over the trend line for the forecast, the regression equation, R-Squared, and p-value are shown. For details about the regression model, right-click on the forecast trend line and select Describe Trend Model.

Lab Exhibit 10.5.8

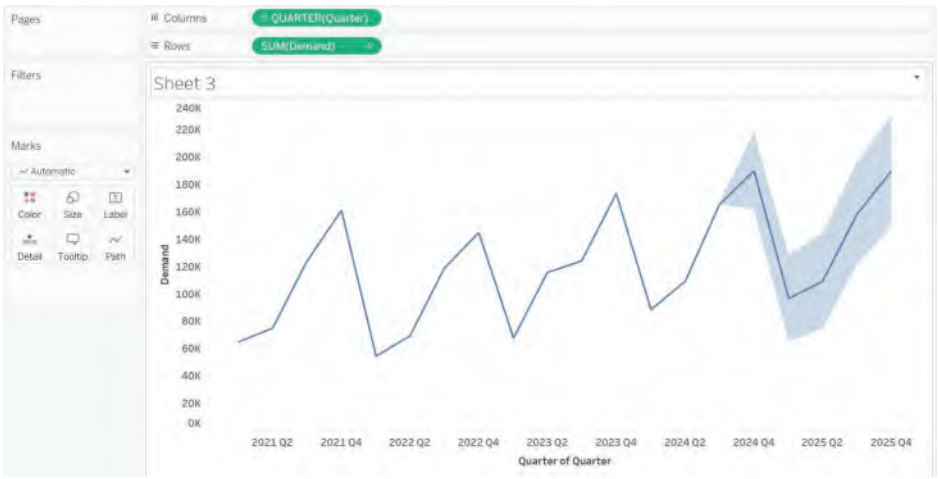
Tableau Software



Analysis Task 5: Make the Forecast Part of the Trend Line.

You have just created a trend line with a forecast in Tableau. However, as you look at the data visualization, you may want the forecasted data points to be part of the original trend line. We can do that by following these steps:

1. To remove the forecasted trend line, drag the Forecast pill out of the Marks card. Note that the original trend line has changed to incorporate the new forecasted data points. The regression model has changed as well (**Analysis > Trend Lines > Describe Trend Model**). The visualization should look like Lab Exhibit 10.5.9. The blue-shaded area is the (default) 95 percent prediction interval for the forecast. That is, there is a 95 percent likelihood that the value of demand will be in the shaded area for the forecast period.



Lab Exhibit 10.5.9

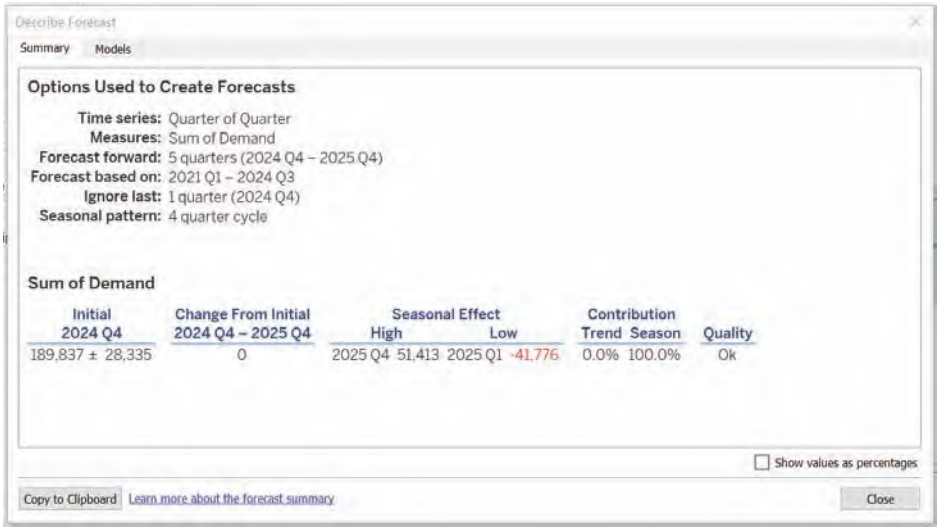
Tableau Software

2. To see details about the forecast, select **Analysis > Forecast > Describe Forecast.** The Describe Forecast dialog box opens. *Note:* The last quarter was ignored, and a seasonal pattern was detected by Tableau. In addition, note that the Quality of the forecast is OK. Tableau assigns a value of GOOD, OK, or POOR to Tableau forecasts. Tableau compares the current forecast model to a naive forecast that assumes the next period value always equals the current period value (or constant demand). GOOD means that the Tableau forecast has less than half of the error of a naive forecast, OK likely has less error, and POOR has more error.[31] Lab Exhibit 10.5.10 shows the Describe Forecast dialog box.

[31]Source: https://help.tableau.com/current/pro/desktop/en-us/forecast_describe.htm. Accessed 7/22/2021.

Lab Exhibit 10.5.10

Tableau Software

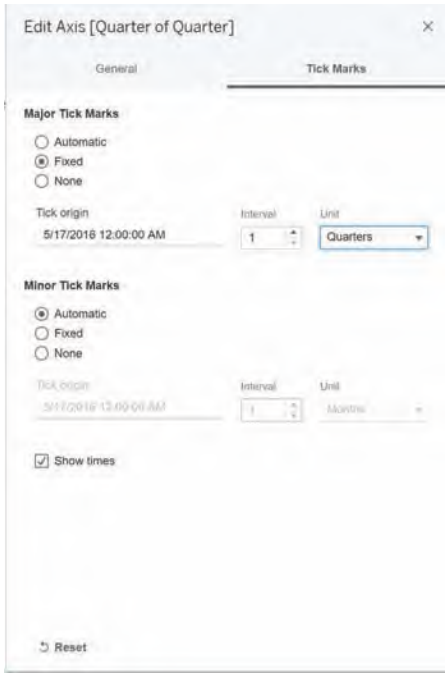


Let's format the visualization by giving it a name, renaming the X-axis and Y-axis, and organizing the X-axis by quarter.

3. To rename Sheet 1, double-click on Sheet 1 and type "1 Year Unit Demand Forecast".
4. Double-click on the X-axis (Quarter of Quarter).
5. The Edit Axis dialog box should appear. In the General tab, change the Axis Title to Quarter (not Quarter of Quarter). Click on the X at the top of the Edit Axis dialog box to save the changes.
6. Repeat the directions in steps 4 and 5 above to change the Y-Axis Title to Demand (Units).
7. Click on the Tick Marks tab (Lab Exhibit 10.5.11).
8. In the Major Tick Marks section, click Fixed.
9. Use the drop-down arrow by Unit to change it to Quarters.
10. Click on the x at the top of the Edit Axis dialog box to save the changes.

Lab Exhibit 10.5.11

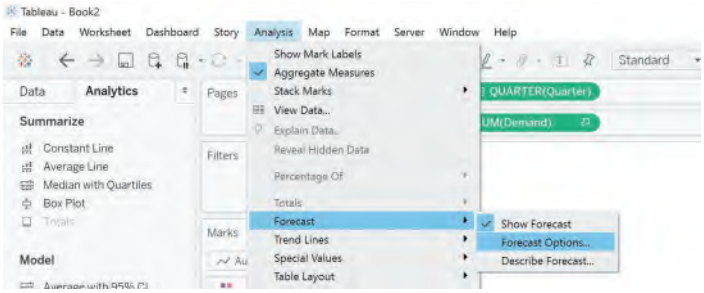
Tableau Software



Analysis Task 6: Change Forecast Assumptions.

Now let's see what happens if we change an underlying assumption of the forecast. The forecast is for one year out, skipping the last quarter's observation. Let's change the forecast options to include the last quarter's data point.

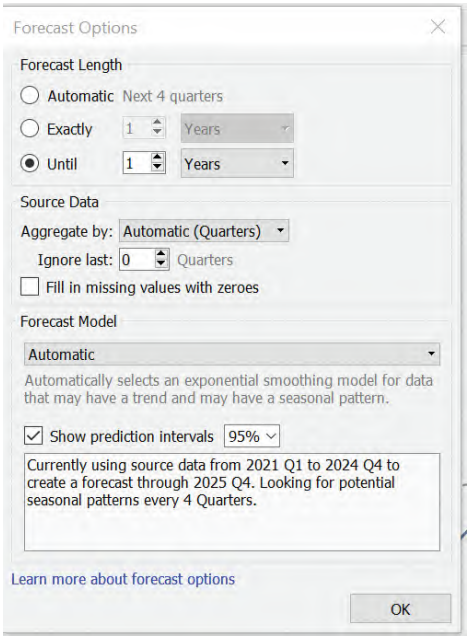
1. Make a copy of the 1 Year Unit Demand Forecast sheet. Right-click the worksheet tab.
2. Select Duplicate. A copy of the first visualization should appear. The sheet will be named 1 Year Unit Demand Forecast (2).
3. From the menu, select **Analysis > Forecast > Forecast Options** as shown in Lab Exhibit 10.5.12.



Lab Exhibit 10.5.12

Tableau Software

4. Under Forecast Length, click Until.
5. Click the up arrow until 1 years appears. Doing so ensures that the forecast goes through 2025 Q4 to match the original forecast.
6. To include all of the historical data in the forecast, in the Source Data section, set Ignore last: to 0. The Forecast Options box should look like Lab Exhibit 10.5.13.



Lab Exhibit 10.5.13

Tableau Software

7. Before we close the Forecast Options dialog box, let's look at the other Forecast Options that you can set. This dialog box allows you to also specify the type of forecast model, indicate how to aggregate the data, fill in missing values with zeros, specify prediction intervals, and choose whether to show prediction intervals. You can explore these options on your own.
8. Click OK. The visualization should change.
9. Rename this new visualization "All Data 1 Year Unit Demand Forecast".

DATA VISUALIZATION

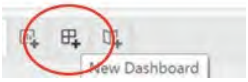
Report the Results: Create a Dashboard

We just created two hoverboard unit product-demand forecasts for one year into the future, each with different underlying assumptions. Don't Fall Inc.'s management can use the forecasts to help plan for the future. To help management easily compare the two forecasts, let's create a dashboard. A dashboard allows multiple data visualizations to be viewed at the same time. Use the following steps to create a dashboard.

1. Click on the New Dashboard tab, as shown in Lab Exhibit 10.5.14.

Lab Exhibit 10.5.14

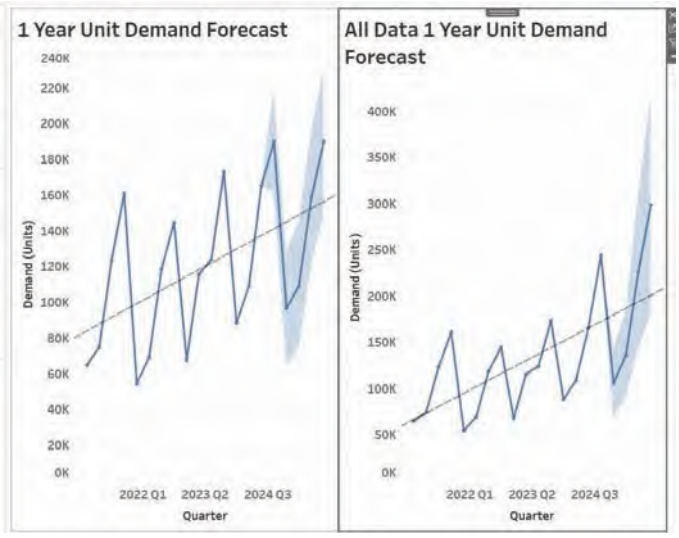
Tableau Software



2. Drag each sheet individually to where Tableau indicates "Drop sheets here". Lab Exhibit 10.5.15 shows what the dashboard should look like. This side-by-side comparison should help management compare the two forecasts.

Lab Exhibit 10.5.15

Tableau Software



The side-by-side comparison allows users to see how the trend line and forecast change when the data set changes. Hovering over each trend line helps users see how the regression line changes as well as how the fit changes.

Assessment

1. Take a screenshot of your dashboard, copy it into a Word document named "Lab 10.5 Tableau Submission.docx", and label the screenshot Submission 1. Take a screenshot of the Describe Trend Model dialog box for the 1 Year Data, paste it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the Describe Forecast dialog box for the 1 Year Data, paste it into the same Word document, and label the screenshot Submission 3. Take a screenshot of the Describe Trend Model dialog box for the All Data, paste it into the same Word document, and label the screenshot Submission 4. Take a screenshot of the Describe Forecast dialog box for the All Data, paste it into the same Word document, and label the screenshot Submission 5.
2. Answer the questions in Connect and upload your Lab 10.5 Tableau Submission.docx via Connect, if assigned.

Alternate Lab 10.5 Tableau

Predictive Analytics/Supply Chain: Forecasting Unit Demand[32]

Apply the same step in Lab 10.5 to the Alt Lab 10.5 Data.xlsx data set.

Required

Create a regression (call a Trend Line in Tableau) and create a forecast for hoverboard unit demand.

Specify the Question

What is the relationship between time and unit demand? What is the expected unit demand for the next year?

Assessment

1. Take a screenshot of your dashboard, copy it into a Word document named "Alt Lab 10.5 Tableau Submission.docx", and label the screenshot Submission 1. Take a screenshot of the Describe Trend Model dialog box for the 1 Year Data, paste it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the Describe Forecast dialog box for the 1 Year Data, paste it into the same Word document, and label the screenshot Submission 3. Take a screenshot of the Describe Trend Model dialog box for the All Data, paste it into the same Word document, and label the screenshot Submission 4. Take a screenshot of the Describe Forecast dialog box for the All Data, paste it into the same Word document, and label the screenshot Submission 5.
2. Answer the questions in Connect and upload your Alt Lab 10.5 submission.docx via Connect, if assigned.

[32]This lab is a very simplified version of an article by Farzan Sajahan, January 18, 2021, "Time Series Analysis in Tableau." Available online at https://dataschoolonline.com/time-series-analysis-in-tableau/. Accessed 7/22/2021. Please see this link for more in-depth information about time series trend lines and forecasts.

Lab 10.6 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Prescriptive Analytics/Human Resources: Optimizing Employee Assignment

Keywords

Predictive Analytics, Excel Solver Add-In, Optimization, Human Resources

Decision-Making Context

Providing actionable intelligence is one goal of using analytics. Prescriptive analytics answers the questions "What should we do, based on what we expect to happen?" and "How do we optimize our performance based on potential constraints?" Prescriptive analytics recommends a course of action.

One prescriptive analytics technique is *optimization,* which uses math to determine the best solution to a problem. The problem could be maximizing something like net income or sales revenues, or minimizing something like costs. In this lab, we use the Excel Solver Add-In, which is designed to solve optimization problems. We will use the Solver Add-In to solve a human resources (HR) problem. Specifically, we will determine the optimal way to assign employees to several tasks with the goal of minimizing labor cost.

The fictitious Eagle Golf Company manufactures different types of customized golf products based on customer specifications. It has developed a new type of golf ball called MYOB (Make Your Own Ball). The balls are transparent and can have anything inside that the customer wants, including flashing lights and 3D team mascots.

To produce the MYOB balls, Eagle Golf Company has three stations that must be staffed by employees. Eagle Golf has hired three new employees for production with varying levels of experience and skills. Therefore, each employee takes a different amount of time to complete the tasks at a workstation. Employees work four-hour shifts.

The constraints are: (1) Each employee can work at one and only one workstation at a time, and (2) each station needs one and only one worker.

Required

1. Determine the optimal employee assignment to minimize labor cost.
2. Create a PowerPoint slide to share your results with Eagle Golf Company's management.

Specify the Question

How should employees be assigned to different tasks to minimize labor cost?

Obtain the Data

The data for this lab are set up slightly different than the data for previous labs (specifically, there are no traditional rows and columns of data). To use the Solver Add-In, it is easiest if the data are set up as in Lab 10.6 Data.xlsx. Below are the definitions of the labels used.

Data Dictionary

Station 1/2/3:	Indicates workstation 1, 2, or 3
Employee 1/2/3:	Indicates employee 1, 2, or 3
Cost Table:	Indicates in dollars how much it will cost Eagle Golf for a specified employee to complete the tasks at a specified station for a four-hour shift; varies based on employee's skill set and experience

Assignment Table:	Indicates employee assignments to a specific station (YES = 1, NO = 0); this is what we are trying to figure out
Station Assigned:	The sum of the work assignments for an employee (= 1, because employees can work at only one station at a time)
Employees Assigned:	The sum of the employee assignment for a station (= 1, because a station can have only one employee assigned to it at a time)
Supply:	Set at 1 for each employee (because employees can work at only one station at a time)
Demand:	Set at 1 for each station (because stations need only one worker at a time)
Total Cost:	Total cost of worker assignments (equal to cost * assignment); thisisthevariablewearetryingtooptimize(orminimizecostinthiscase)

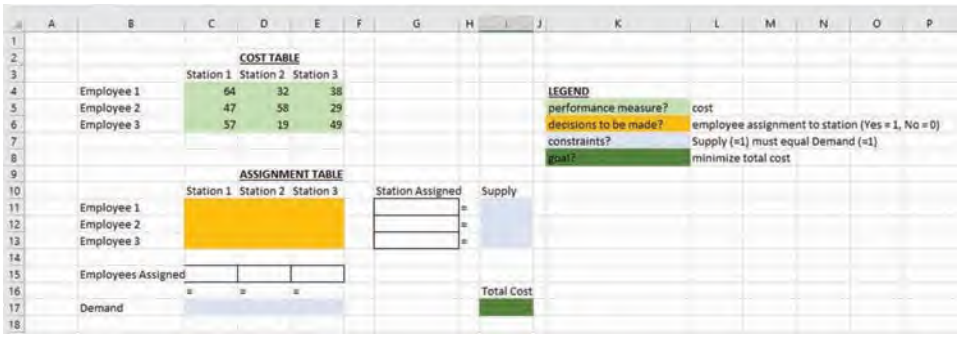
For ease of use, the following ranges have been named as follows. To see the name in the Name Box, select the cells as shown in Lab Exhibit 10.6.1, which will be helpful when we use the Solver.

Lab Exhibit 10.6.1

Range Name	Cells
Cost	C4:E6
Assignment	C11:E13
EmployeesAssigned	C15:E15
Demand	C17:E17
StationAssigned	G11:G13
Supply	I11:I13
TotalCost	I17

Lab Exhibit 10.6.2 shows the spreadsheet before we complete any steps.

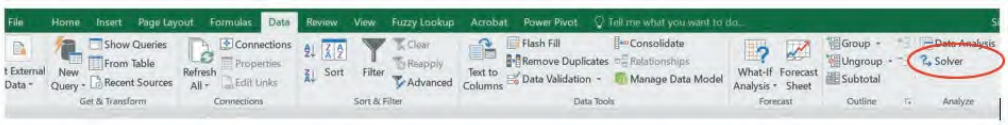
Lab Exhibit 10.6.2



Analyze the Data

Analysis Task 1: Install the Solver Add-In.

1. The first step is to install the Solver Add-In if it is not already installed. If Solver is installed, it will appear in the **Data > Analyze** menu option, as shown in Lab Exhibit 10.6.3. If Solver does not appear, the textbook Appendix F provides instructions for installing it.



Lab Exhibit 10.6.3

Analysis Task 2: Set Up the Worksheet.

1. Using the information in Lab Exhibit 10.6.4, enter the formulas and functions in Station Assigned, Employees Assigned, and Total Cost; set both Supply and Demand equal to all 1s because each station needs one, and only one, employee, and each employee can work at one, and only one, station. Note the formula for Total Cost = SUMPRODUCT(Cost,Assignment). SUMPRODUCT is an Excel function that multiples values from the first set of cells (Cost) with values from the second set of cells (Assignment) and adds them up.

Lab Exhibit 10.6.4

	A	B	C	D	E	F	G	H	I
1									
2				COST TABLE					
3			Station 1	Station 2	Station 3				
4		Employee 1	64	32	38				
5		Employee 2	47	58	29				
6		Employee 3	57	19	49				
7									
8									
9				ASSIGNMENT TABLE					
10			Station 1	Station 2	Station 3		Station Assigned		Supply
11		Employee 1					=SUM(C11:E11)	=	1
12		Employee 2					=SUM(C12:E12)	=	1
13		Employee 3					=SUM(C13:E13)	=	1
14									
15		Employees Assigned	=SUM(C11:C13)	=SUM(D11:D13)	=SUM(E11:E13)				
16			=	=	=				Total Cost
17		Demand	1	1	1				=SUMPRODUCT(Cost,Assignment)

Analysis Task 3: Solve via the Manual Approach.

At this point we could try trial and error to determine the lowest Total Cost. In other words, we could assign workers to different stations using 0s and 1s in cells C11 through E13. Lab Exhibit 10.6.5 calculates a Total Cost of \$171 = (1*64) + (1* 58) + (1*49) when Employee 1 is assigned to Station 1, Employee 2 is assigned to Station 2, and Employee 3 is assigned to Station 3. To determine the cost, you would manually have to try every possible combination, which is not too difficult for three employees and three stations, but imagine if there were hundreds of employees and hundreds of stations!

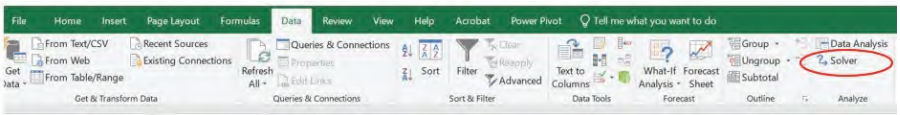
Lab Exhibit 10.6.5

	A	B	C	D	E	F	G	H	I
1									
2				COST TABLE					
3			Station 1	Station 2	Station 3				
4		Employee 1	64	32	38				
5		Employee 2	47	58	29				
6		Employee 3	57	19	49				
7									
8									
9				ASSIGNMENT TABLE					
10			Station 1	Station 2	Station 3		Station As		Supply
11		Employee 1	1	0	0		1	=	1
12		Employee 2	0	1	0		1	=	1
13		Employee 3	0	0	1		1	=	1
14									
15		Employees Assigr	1	1	1				
16			=	=	=				Total Cost
17		Demand	1	1	1				171

Analysis Task 4: Solve via Solver.

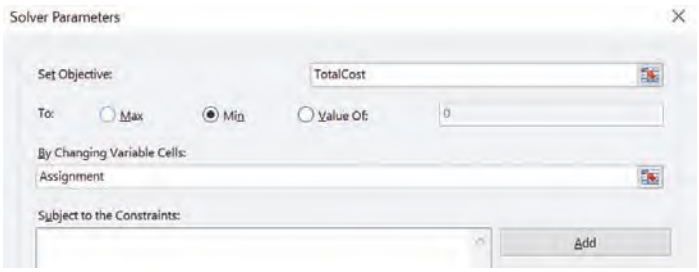
Instead of using trial and error to determine the best assignments to minimize cost, we can use the Excel Solver to calculate the optimal answer.

1. From the menu, select **Data > Analyze > Solver** as shown in Lab Exhibit 10.6.6.



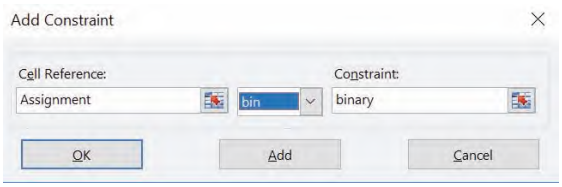
Lab Exhibit 10.6.6

2. The Solver Parameters dialog box should appear. We will use the table names for the predefined ranges of cells, as shown in Lab Exhibit 10.6.7. Enter "TotalCost" as the Set Objective.
3. Click on Min.
4. Enter "Assignment" in the By Changing Variable Cells: as shown in Lab Exhibit 10.6.7. These parameters indicate to Excel that the TotalCost cell (the Set Objective) needs to be minimized by changing the assignments in cells C11:E13, which is named Assignment in Lab Exhibit 10.6.7. This setting directs Excel to determine the solution with the smallest TotalCost.



Lab Exhibit 10.6.7

5. To complete the Subject to the Constraints field, click Add.
6. The Add Constraint dialog box appears. Enter "Assignment" as the Cell Reference.
7. Select "bin" from the middle drop-down.
8. Enter "binary" as the Constraint. These parameter choices indicate to Excel that Assignment can only be a "0" or a "1," or one employee can work at a station or not. Your Add Constraint dialog box should look like Lab Exhibit 10.6.8.



Lab Exhibit 10.6.8

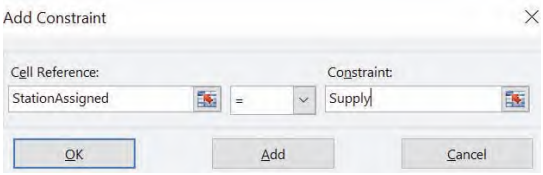
9. Click OK.
10. Click Add again.
11. In the Add Constraint dialog box, enter "EmployeesAssigned" as the Cell Reference.
12. Select "=" from the middle drop-down.
13. Enter "Demand" as the Constraint. These parameter choices indicate to Excel that Employees Assigned must equal the Demand for a station (which has been set to 1). Your Add Constraint dialog box should look like Lab Exhibit 10.6.9.

Lab Exhibit 10.6.9



14. Click OK.
15. Click Add again.
16. In the Add Constraint dialog box, enter "StationAssigned" as the Cell Reference.
17. Select "=" from the middle drop-down.
18. Enter "Supply" as the Constraint. These parameter choices indicate to Excel that Station Assigned must equal the Supply for a station (which has been set to 1). Your Add Constraint dialog box should look like Lab Exhibit 10.6.10.

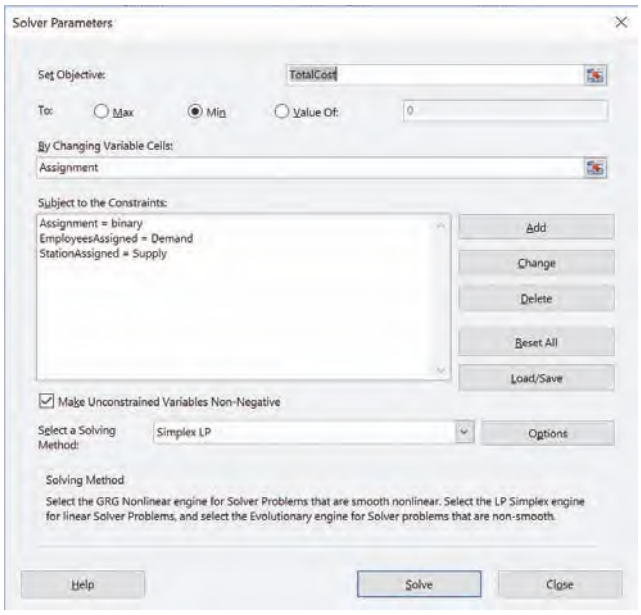
Lab Exhibit 10.6.10



19. Click OK.
20. You should be returned to the Solver Parameters dialog box. Check Make Unconstrained Variables Non-Negative. When this box is checked, Excel sets the lower bound for the decision variables cells to zero unless an explicit lower bound has been set in the constraints.
21. For Select a Solving Method, use the down arrow to select Simplex LP. This tells Excel which method to use to optimize the answer. Simplex LP is a robust method for determining the globally optimal solution for linear functions. Lab Exhibit 10.6.11 shows what the Solver Parameters dialog box should look like.

Linear function: An equation that creates a straight line on a graph. Its formula is: y = mx + b.

Lab Exhibit 10.6.11



22. Click Solve.
23. Click OK when the Solver Results dialog box appears. The optimal assignment results will appear in cells C11:E13, along with the total cost in cell I17. If Solver does not find a result, double-check all of the formulas for errors.

Report the Results

We use Excel Solver to determine the optimal assignment of employees to minimize cost. Eagle Golf Company can use this information to optimally assign employees to the different stations. Create a PowerPoint Slide to show the optimal assignment and total cost for a four-hour shift.

Assessment

1. Take a screenshot of A1:I17, copy it into a Word document named "Lab 10.6 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of your PowerPoint slide, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Lab 10.6 Excel Submission.docx via Connect, if assigned.

Alternate Lab 10.6 Excel

Prescriptive Analytics/Human Resources: Optimizing Employee Assignment

Apply the same steps in Lab 10.6 to the Alt Lab 10.6 Data.xlsx data set.

Required

1. Determine the optimal employee assignment to minimize labor cost.
2. Create a PowerPoint slide to share your results with Eagle Golf Management.

Specify the Question

How should employees be assigned to different tasks to minimize labor cost?

Assessment

1. Take a screenshot of A1:I17, copy it into a Word document named "Alt Lab 10.6 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of your PowerPoint slide, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Alt Lab 10.6 Excel Submission.docx via Connect, if assigned.

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Chapter 11

Advanced Business Analytics

A Look Back

- Chapters 7 through 10 described how companies use business analytics and the SOAR model to gain actionable business intelligence (BI) across four main business functions: marketing, accounting, finance, and operations.
- These chapters focused on the first four types of analytics: descriptive, diagnostic, predictive, and prescriptive.
- The last chapter, Chapter 10, focused on operations, discussing examples of operations questions addressable using business analytics, the sources of operations data, key types of operations analytics, and the reporting of the results of operations analytics.

A Look at This Chapter

- Advances in technology are being applied to business analytics to create a fifth type of analytics, adaptive/autonomous analytics, which mimics human capabilities, such as judgment and reasoning, to create insights from structured and unstructured data.
- Adaptive/autonomous analytics in combination with the SOAR model takes analytics to a new level as reiterative technology executes decisions, scans the environment, adapts to changes, and then executes new decisions.
- The technology underlying adaptive/autonomous analytics includes artificial intelligence and machine learning. Companies can pull everything together using robotic process automation and cockpits.

A Look Ahead

- Chapter 12 asks you to use the SOAR model to answer a specific business-related question. It will help you synthesize all of the concepts covered in this textbook.



Waymo Robocar
Sundry Photography/Shutterstock

Driverless Driving

"With every mile we drive, we never stop learning."
—WAYMO

Can you imagine a world in which cars drive themselves? Autonomous or self-driving cars, also known as robocars, are quickly becoming a reality. Beginning in 2019, **Waymo**, originally started by **Google** in 2009 as a self-driving car project, began transporting passengers on California roads as part of California's Autonomous Vehicle Passenger Service pilot. Waymo cars provide free rides to California passengers with a safety (human) driver behind the wheel. After a similar pilot in Phoenix, Arizona, Waymo recently launched Waymo One, which works just like **Uber** or **Lyft**. A Waymo One robotaxi with a safety driver behind the wheel (at least for now) picks passengers up and drives them to their destination. In 2021, California began issuing driverless permits to Waymo and several other robocar developers (including a General Motors–backed competitor, **Cruise**), meaning that a safety driver is no longer in the car.

An enormous amount of technology makes robocars possible. First, roads must be mapped using three-dimensional maps that capture not only the road but also curbs, driveways, stop signs, cross walks, and other permanent structures. The robocars use these maps to know exactly where they are. In addition, cars are equipped with lasers, cameras, radars, sonar, and GPS systems that scan the road ahead as far as three football fields in every direction to detect objects, traffic controls, and road conditions. Software predicts the movements of all objects and then determines how the car should react. This chapter helps you understand how companies can utilize the technology used in robocars to help drive their businesses.

Sources: Kirsten Korosec, Waymo is now allowed to transport passengers in its self-driving vehicles on California roads, https://techcrunch.com/2019/07/02/waymo-is-now-allowed-to-transport-passengers-in-its-self-driving-vehicles-on-california-roads/July 2, 2019; Kirsten Korosec, Waymo launches self-driving car service Waymo One, https://techcrunch.com/2018/12/05/waymo-launches-self-driving-car-service-waymo-one/; December 5, 2018; https://waymo.com; Companies website

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

- **LO 11.1** Explain adaptive/autonomous analytics and its relationship to other types of analytics.
- **LO 11.2** Describe advanced analytics methods.
- **LO 11.3** Give examples of advanced analytics used in the various business functions.
- **LO 11.4** Describe trending technologies.
- **LO 11.5** Describe data-driven organizations and the future of business analytics.

LABS FOR CHAPTER 11

- **LAB 11.1** **Excel:** Advanced Analytics/Marketing: Analyzing Customer Repurchase Rates Using Cohorts
- **LAB 11.2** **Tableau:** Advanced Analytics/Marketing: Analyzing Text Using a Word Cloud
- **LAB 11.3** **Excel:** Advanced Analytics/Human Resources: Visualizing Survival Analysis
- **LAB 11.4** **Excel:** Advanced Analytics/Supply Chain: Estimating Demand Using Monte Carlo Simulation
- **LAB 11.5** **Advanced Analytics:** Understanding Blockchain

LO 11.1

Explain adaptive/autonomous analytics and its relationship to other types of analytics.

11.1 ADAPTIVE/AUTONOMOUS BUSINESS ANALYTICS

"May you have the hindsight to know where you've been, The foresight to know where you are going, And the insight to know when you have gone too far."

—IRISH BLESSING

Earlier chapters in this book focused on four types of analytics that answer different questions. Specifically:

- Descriptive analytics answers the question "What happened in the past?"
- Diagnostic analytics answers the question "Why did it happen (in the past)?"
- Predictive analytics answers the question "What is likely to happen in the future?"
- Prescriptive analytics answers the question "What action(s) should we take?"

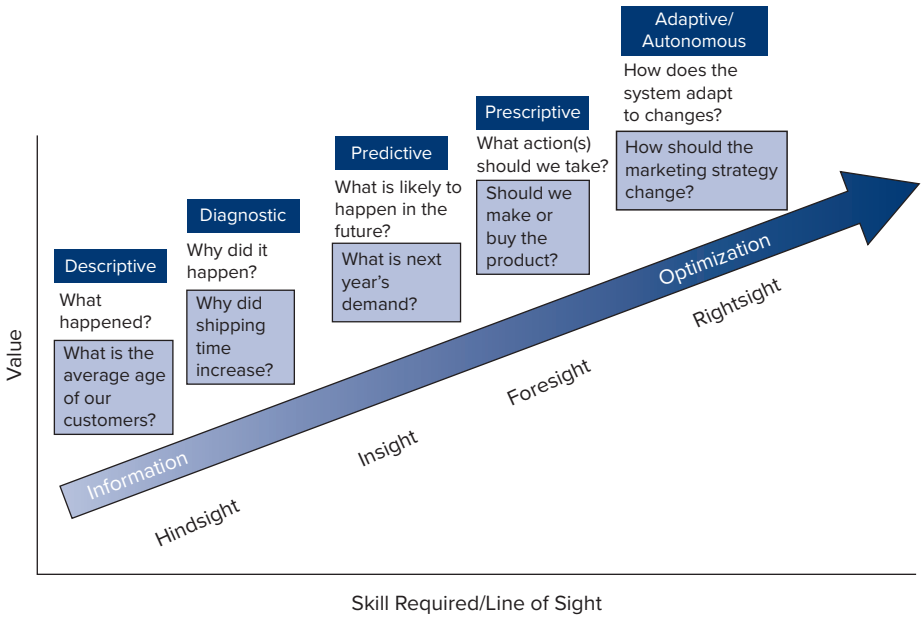
The four types of analytics are interrelated and typically build on each other. In other words, to generate actionable intelligence with prescriptive analytics, you must understand the organizational data, which means that you must be able to understand the results from lower levels of analytics—that is, descriptive, diagnostic, and predictive analytics.

Chapter 1 introduced the analytics ascendancy model in Exhibit 1.6, repeated here in Exhibit 11.1. The analytics ascendancy model shows that as we move from descriptive to prescriptive analytics, two things happen simultaneously: (1) the difficulty of the analysis increases, requiring more sophisticated tools, and (2) the value of the actionable intelligence increases. The value increases because the company gains more "sight" from the data:

- *Hindsight* comes from using descriptive analytics to generate historic reports.
- *Insight* comes from using diagnostic analytics to create real-time (or dynamic) reports.
- *Foresight* comes from using predictive and prescriptive analytics to predict outcomes.

Exhibit 11.1 Analytics Ascendancy Model

Source: Adapted from: Information Builders, 2018. Raef Lawson, "New Competencies for Management Accounts," *Strategic Finance*, March 2019, https://sfmaganize.com/post-entry/march-2019-new-competencies-for-management-accountants/ 8/9/2022. Accessed 8/9/2022



While the first four types of analytics can provide hindsight, insight, and foresight into business questions, adaptive/autonomous analytics provides a vehicle to propel a company forward with *rightsight,* as described in the next section.

ADAPTIVE/AUTONOMOUS ANALYTICS: MIMICKING HUMAN CAPABILITIES

In recent years, technological advances have allowed a new type of analytics to emerge: **adaptive/autonomous analytics**, also called **cognitive analytics**. This new type of analytics mimics human capabilities, such as judgment and reasoning, creating insights from structured and unstructured data. Unlike current predictive systems, which are based on human-coded **algorithms**, or sets of programmed rules, adaptive/autonomous analytics consists of algorithms that continuously learn and rapidly modify themselves to optimize performance as new data and outcomes are created.

Adaptive/autonomous analytics takes business analytics to a new level as reiterative technology executes decisions, scans the environment, adapts to changes, and then executes new decisions. Essentially, it answers the questions "How does the system adapt to changes?" and "How can we continuously run analytic solutions?" It provides the company with *rightsight* by "learning the right question to ask your data, at the right time, at the right place, for the right object (or event), in the right context"[1] so that actionable intelligence can be delivered to the right people in the right form at the right time.

How does adaptive/autonomous analytics deliver more value to a company than the other analytics types? It can automate, combine, and optimize technologies used in the other four types of analytics and then add additional technologies to enhance analytics even more. In addition, adaptive/autonomous analytics is not limited by human capabilities, fatigue, and slow processing speed. It allows humans to be more creative, consider more options, and tackle new and more complex tasks and decisions as the technology handles much of the "grunt work," freeing us from personal biases or beliefs that may impact the data sets we analyze and how we analyze them. In addition, operating multiple adaptive/autonomous systems in parallel can help synchronize processes to create an optimized, autonomous operational system. Together, all of these capabilities create a reiterative SOAR model, helping companies make better decisions that increase productivity and quality, decrease costs, and increase revenues.

A good example of adaptive/autonomous analytics is the robocar described at the beginning of this chapter. It is impossible to create human-programmed, rule-based algorithms for every possible scenario that a driver encounters on the road. Instead, robocars are trained to drive themselves based on billions of hours of real and simulated practice. **Waymo's** primary tag line is "We're Building the World's Most Experienced Driver," as the training is equivalent to hundreds of years of human driving experience. Exhibit 11.2 shows some of the features of a robocar.

Robocars run on data collected from the car itself as well as data stored in the cloud about weather, traffic, GPS location, and other robocars. In other words, robocars depend on Big Data. Real-time adaptive/autonomous analytics uses the data to make many driving decisions, including braking, avoiding collisions, recognizing traffic signs and signals, and even predicting what human drivers in other cars will do. Exhibit 11.3 shows how all of this technology works together.

[1]Source: Kirk Borne, September 9, 2021 "Cognitive Analytics Answers the Question: What's Interesting in Your Data." Available online at https://medium.com/@kirk.borne/cognitive-analytics-answers-the-question-whats-interesting-in-your-data-3406a0e09cf9. Accessed 11/15/2021.

Exhibit 11.2
Robocar Features

Source: Raoul Ranoa. https://www.trbimg.com/img-5386922b/turbine/la-sci-g-google-self-driving-car-20140528. Accessed on 11/15/2021.

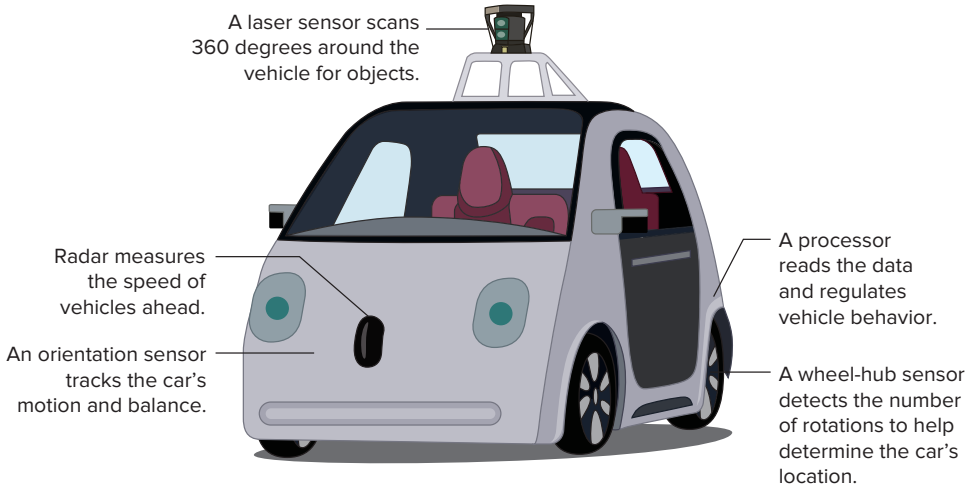
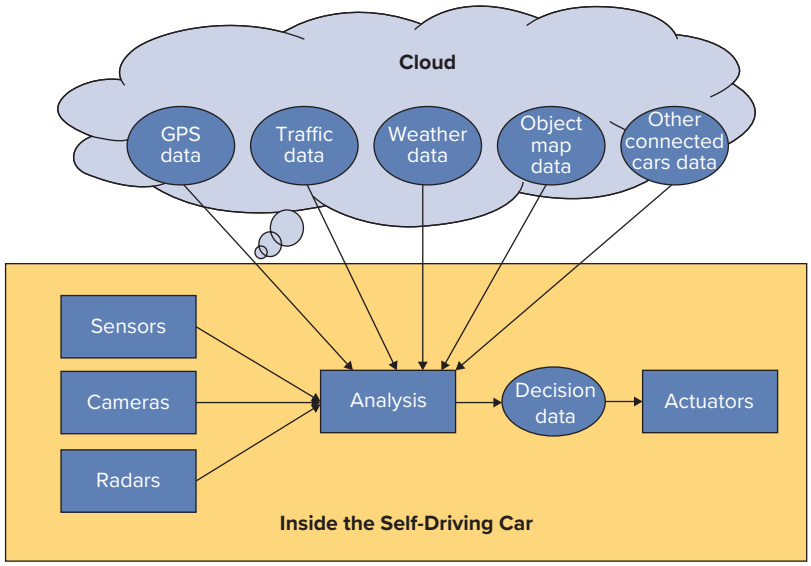


Exhibit 11.3
Robocar Inner Workings

Source: Ria Cheruvu. 2015. "Big Data Applications in Self-Driving Cars." Working paper. Available online at https://www.researchgate.net/publication/318940413_Big_Data_Applications_in_Self-Driving_Cars. Accessed 11/15/2021.



Once the technology is mastered, robocars are projected to:[2]

- reduce the number of car accidents because, unlike human beings, fully functioning robocars do not get distracted and are better at anticipating problems than humans are.
- decrease traffic congestion and at the same time allow more cars on the road (because robocars space automatically).
- reduce gas consumption due to constant travel speeds and lighter cars.
- increase human productivity by allowing people to work rather than spend their time driving or parking a car.
- increase mobility for people who are unable to drive.
- reduce traffic enforcement personnel because robocars don't speed.

[2]Lyndon Bell, "10 Benefits of Self-Driving Cars." Available online at https://www.autobytel.com/car-ownership/advice/10-benefits-of-self-driving-cars-121032/. Accessed 8/08/2019.

How are forward-thinking companies planning to use self-driving car technology to expand their product lines, improve safety, and reduce costs? Consider these examples:

- **UPS** has been testing robotrucks between Tucson and Phoenix, Arizona. These robotrucks feature lane-departure technology, advanced braking systems, and sensor technology. Self-driving trucks may soon outnumber self-driving taxis, because delivery routes are well defined, and driverless long-haul trucks can run at night when traffic is lighter and more predictable.[3]
- **Walmart** is using driverless robotrucks to deliver groceries from a Walmart fulfillment center to a Walmart neighborhood grocery store in Bentonville, Arkansas.[4]
- **FedEx** is testing robotrucks between Dallas and Houston, Texas, with a driver (at least for now). The company hopes to use these robotrucks to compensate for driver shortages in the future.[5]

Here are some examples of other companies' use of adaptive/autonomous analytics.

- **Charles Schwab's** Intelligent Portfolios, also known as robo-advisors, use adaptive/autonomous analytics to actively manage customers' investments based on their responses to a questionnaire on goals, time horizon, and risk preferences.[6]
- **Deloitte** uses adaptive/autonomous analytics to read legal contracts to determine risks and proper accounting treatment.[7]
- **Five Guys Burgers and Fries** uses adaptive/autonomous analytics to help screen, review, and interview job candidates.[8]
- **Ford** uses adaptive/autonomous analytics for quality assurance, including detecting wrinkles in car seats.[9]
- **Unilever** uses adaptive/autonomous analytics to optimize output, use material optimally, and reduce energy costs in a manufacturing plant that makes ice cream and Dove soap.[10]

[3]Source: Reuters, 2019, "UPS Buys Stake in TuSimple After Using Its Autonomous Trucks." Available online at https://www.autoblog.com/2019/08/16/ups-tusimple-self-driving-trucks/. Accessed 8/26/2019.
[4]Source: S. Dent, November 8, 2021, "Walmart Is Using Driverless Trucks to Complete a 7-mile Delivery Loop," *Engadget.* Available online at https://www.engadget.com/walmart-is-using-driverless-trucks-to-complete-a-7-mile-delivery-loop-143016425.html. Accessed 11/15/2021.
[5]Ihsaan Fanusie, October 3, 2021, "Aurora, FedEx Team Up on Driverless Tech to Address Trucker Shortage." *Yahoo! Finance.* Available online at https://www.msn.com/en-us/money/companies/aurora-fedex-team-up-on-driverless-tech-to-address-trucker-shortage/ar-AAP5O7x. Accessed 11/15/2021.
[6]Charles Schwab, 2021, *Schwab Intelligent Portfolios.* Available online at https://intelligent.schwab.com/. Accessed 11/15/2021.
[7]Source: Deloitte, 2022, "Introducing dTrax—Deloitte's Artificial Intelligence (AI) Contract Analysis Solution." Available online at https://www2.deloitte.com/us/en/pages/advisory/solutions/dtrax-ai-contract-analysis-solution.html. Accessed 5/17/2022.
[8]Source: Amy Elisa Jackson, December 31, 2018, "Popular Companies Using AI to Interview and Hire You." *Glassdoor.* 2018. Available online at https://www.glassdoor.com/blog/popular-companies-using-ai-to-interview-hire-you/. Accessed 11/15/2021.
[9]Bernard Marr, May 17, 2019, "The Amazing Ways the Ford Motor Company Uses Artificial Intelligence and Machine Learning," *Forbes.* Available online at https://www.forbes.com/sites/bernardmarr/2019/05/17/the-amazing-ways-the-ford-motor-company-uses-artificial-intelligence-and-machine-learning/?sh=6ccfd1be49a2. Accessed 11/15/2021.
[10]Jennifer Smith, July 15, 2019, "Unilever Uses Virtual Factories to Tune-up Its Supply Chain," *Wall Street Journal.* Available online at https://www.wsj.com/articles/unilever-uses-virtual-factories-to-tune-up-its-supply-chain-11563206402. Accessed 8/08/2022.

PROGRESS CHECK

1. In the context of analytics, what does adaptive mean? What does autonomous mean?
2. Describe how companies can use adaptive/autonomous analytics to customize products for customers. For example, how could **MetLife** use adaptive/autonomous analytics to help insurance customer service representatives provide better customer service via the phone?

LO 11.2

Describe advanced analytics methods.

11.2 ADVANCED ANALYTIC METHODS

Advanced analytics methods make adaptive/autonomous analytics possible. In this section, we explore the primary advanced analytics methods in use today.

ARTIFICIAL INTELLIGENCE

Artificial intelligence, or **AI,** is "a collective term for computer systems that can sense their environment, think, learn, and take action in response to what they're sensing and their objectives."[11] AI systems accomplish tasks in a way that appears smart or "intelligent" to us as they appear to mimic human decision processes. Some AI systems interact with, or assist, humans with making decisions and completing tasks. Others don't. Some AI systems adapt to, or learn, from their environment. Others don't.

Based on the business question and/or need, companies determine the best type of AI to use. Exhibit 11.4 categorizes the types of AI in use today. One example of *assisted intelligence* (upper-left quadrant of Exhibit 11.4) is **Google**'s Gmail sorting email into primary, social, and promotion tabs. **Netflix**'s content suggestions are an example of *augmented intelligence* because they provide recommendations based on a person's prior video selections. Automating the re-ordering of inventory is an example *of automation of existing tasks.* The bottom-right quadrant represents the current pinnacle of business analytics: *autonomous intelligence.* A robocar is a good example of task-specific autonomous intelligence. When the robocar can shop for you and cook your dinner, then the technology will have moved closer to true generalized autonomous intelligence.

	Assist Humans with Decision-Making	No Human Involved
Hardwired systems that do not learn	Assisted Intelligence	Automation of existing tasks
Adaptive systems that learn	Augmented Intelligence	Autonomous Intelligence • task-specific, applied, narrow, or weak • general purpose, generalized, strong, or human-level

Exhibit 11.4
Forms of AI in Use Today

Source: PwC, 2017, "Sizing the Price: What's the Real Value of AI for Your Business and How Can You Capitalise?" Available online at https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html. Accessed 11/15/2021.

[11]PwC, 2017, "Sizing the Price: What's the Real Value of AI for Your Business and How Can You Capitalise?" Available online at https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html. Accessed 11/15/2021.

While AI applications have only a modest impact now, **PwC** estimates that in 2030 AI may contribute $15.7 trillion to the world economy, making it "the biggest commercial opportunity in today's fast-changing economy." Value will be created through increased productivity, personalization of products, saved time, and improved product quality.[12]

It is important to understand that AI is not just one type of technology. Rather, it is a catchall phrase for a wide variety of techniques used to process data in a "thinking" manner. We describe two different types of AI in the following sections: machine learning and artificial neural networks.

MACHINE LEARNING

While early applications of AI focused on solving complex calculations, many AI applications now focus on machine learning. **Machine learning (ML)** uses statistical techniques that allow the computer to "learn" from data without being explicitly programmed. Essentially, machine learning automates analytical model building by learning from large data sets, identifying patterns in the data, and then independently making decisions.[13]

Machine learning sometimes uses **data mining**, the process of analyzing large amounts of data to identify patterns, anomalies, and relationships. Data mining may merge data from a variety of internal and external sources. Internal data sources include the enterprise resource planning systems, **data warehouses** (for structured data), and **data lakes** (for both structured and unstructured data). (For a review, see Chapter 2.) External data sources include websites; weather data and forecasts; and government data, including the Consumer Price Index, GDP, and economic forecasts.

Data mining can be used to answer questions like these:

- "Are machine breakdowns related to specific events?"
- "Do some shoppers who are buying diapers also buy beer?"
- "Is weather related to the sale of certain products?"

Data mining generates rules about the data or the relationships/connections in the data, and then machine learning teaches the computer to understand those rules.[14] For example, as an **IBM** commercial says, European bakery customers buy more cake when it rains, but more paninis/sandwiches when it is sunny.[15] Changing production to match the weather helped some European bakeries eliminate waste, increasing their profitability by up to 20 percent.

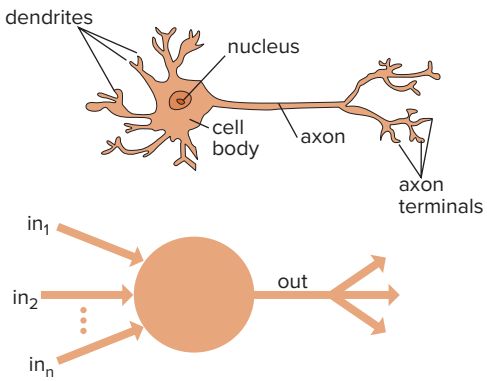
ARTIFICIAL NEURAL NETWORKS

Most AI systems today are *applied*, *narrow*, or *weak* AI systems designed for a specific task. For example, a computer designed to beat a human player at chess may win at chess, but it is unable to do anything else. Robocars combine a series of applied AI applications. However, a long-term goal is to create a general-purpose AI that is *generalized*, *strong*, or *human-level*. This advanced form of AI will be able to function in multiple environments just as humans do.

[12]Information in the paragraph is based on PwC, 2017, "Sizing the Price: What's the Real Value of AI for Your Business and How Can You Capitalise?" Available online at https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html. Accessed 11/15/2021.
[13]From https://www.sas.com/en_us/insights/analytics/machine-learning.html.
[14]From "Data Mining vs Machine Learning: What's the Difference?" Available online at https://www.import.io/post/data-mining-machine-learning-difference/. Accessed 8/21/2019.
[15]See https://www.ispot.tv/ad/7VBe/ibm-bakery. Accessed 4/29/2020.

To achieve this human-level, general AI, scientists are creating **artificial neural networks (ANNs)** to mimic human intelligence. ANNs are one way to implement machine learning. Exhibit 11.5 shows the similarities between one human brain neuron and one ANN processing node. Like the human brain, which contains billions of connected neurons that transfer and process information using electrical signals, an ANN consists of thousands or millions of interconnected processing nodes. The nodes are organized into different layers. The bottom layer is the input layer. The middle layers transform the data based on weights and thresholds that are derived from training data. The top layer is the output layer.

Exhibit 11.5
The Human Brain versus Artificial Neural Networks



Source: Dishashree Gupta, January 30 2020, "Fundamentals of Deep Learning—Activation Functions and When to Use Them?," Available online at https://www.analyticsvidhya.com/blog/2020/01/fundamentals-deep-learning-activation-functions-when-to-use-them/. Accessed 11/15/2021.

Simple ANNs were first developed in the 1950s. They consisted of only one layer of nodes due to technological limitations. In contrast, some of today's ANNs are sometimes 50 layers deep, with the additional layers helping to better classify/categorize the input. ANNs with more than three neural network layers are called **deep learning**.[16] Deep learning can be trained to recognize, classify, and describe unstructured data such as text, images, and voice. For example, by providing multiple photos that are labeled by name and/or category (baby, teenager, adult), deep learning may learn to identify a person (or a category of people) based on pictures by understanding the patterns in the data. The more training examples provided, the better the accuracy of the prediction. Companies are now using ANNs to:

- segment a market on the basis of multiple factors (target marketing)
- predict changes in the stock market
- approve loans
- segment the insurance market to determine insurance rates
- identify fraudulent insurance claims
- schedule projects, machinery, and assembly lines
- identify defective products
- identify shoplifters

[16]This section is based on the information from Larry Hardesty, April 14, 2018, "Explained: Neural Networks." *MIT News Office.* Available online at https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414. Accessed 11/15/2021.

ADVANCED ANALYTICS AND YOUR CAREER

One job in adaptive/autonomous analytics is a cognitive consultant. Cognitive consultants help companies implement cognitive, or AI, technologies to enhance products and processes. In doing so, they help their company gain insights to improve different aspects of the business, such as costs, revenues, or service levels.[17] This career is in hyper-growth mode, with cognitive consultants earning an average salary of $65,263 and top earners making more than $147,500.[18] **IBM** employs cognitive consultants in its Watson Cognitive Solutions and Services group, which uses IBM Watson to leverage unstructured data to drive efficiency and reinvent clients' strategy and operations. To land a job in cognitive consulting, you must have knowledge of advanced analytics techniques, natural language processing, and machine learning.

SOARing to Success

Fun Fact: The History of IBM's Watson

IBM Watson is named for the founder and first CEO of IBM, Thomas J. Watson. In 2006, IBM Research began creating IBM Watson, a computer with software designed to find answers more efficiently than a typical internet search engine. Initially, the software was fed information from a variety of sources, including World Book Encyclopedia, Wikipedia, and Project Gutenberg books.[19] Today, it understands human language; accesses and evaluates data from a wide variety of sources, including social media, the cloud, and Internet of Things; and adapts and learns from user responses. In its public debut in 2011, it beat *Jeopardy's* two biggest all-time champions. Since then, IBM Watson has been expanded and commercialized for healthcare, financial services, and other industries.

PROGRESS CHECK

3. Which of these terms refers to the broadest set of advanced analytics methods: artificial intelligence, machine learning, or artificial neural networks?
4. How could **Facebook** use deep learning? For example, what information could Facebook potentially determine from posted pictures?

11.3 ADAPTIVE/AUTONOMOUS ANALYTICS IN THE BUSINESS FUNCTIONS[20]

LO 11.3

Give examples of advanced analytics used in the various business functions.

Chapters 7–10 discussed how descriptive, diagnostic, predictive, and prescriptive analytics can be used across business functions. Just as these four types of analytics work in the SOAR model, so does adaptive/autonomous analytics. While adaptive/autonomous

[17]Source: David Schatsky, Craig Muraskin, and Ragu Gurumurthy, "Cognitive Technologies: The Real Opportunities for Business." *Deloitte Review* Issue 16. Available online at https://www2.deloitte.com/us/en/insights/deloitte-review/issue-16/cognitive-technologies-business-applications.html. Accessed 12/6/2021.
[18]Source: ZipRecruiter. "Cognitive Consulting Salary." Available online at https://www.ziprecruiter.com/Salaries/Cognitive-Consulting-Salary. Accessed 11/15/2021.
[19]Source: IBM, "A Computer called Watson." Available online at https://www.ibm.com/ibm/history/ibm100/us/en/icons/watson/. Accessed 12/6/2021.
[20]Information in this section is based on PwC, 2017, "Sizing the Price: What's the Real Value of AI for Your Business and How Can You Capitalise?" Available online at https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html. Accessed 11/15/2021.

analytics is just beginning to be used, **PwC** has identified almost 300 opportunities to use AI in a variety of industries and business functions. Following the organization of Chapters 7–10, we now discuss a few examples of AI in marketing, accounting, finance, and operations. Each example may have applications in other business functions.

MARKETING ANALYTICS

"Predictive analytics, as well as prescriptive analytics, are powered by artificial intelligence and machine learning. AI and machine learning hold the promise of providing unparalleled levels of business intelligence and actionable insights to brands."

—**MIKE KAPUT**, Marketing Artificial Intelligence Institute[21]

Marketing analytics is a great place for a company to start exploring adaptive/autonomous analytics. You may not realize it, but you have already used at least one type of advanced marketing analytics: **recommendation engines**, which are data-filtering tools that use algorithms to predict what a user may or may not like. If you have ever viewed a product on a retailer's website and then received recommendations for other products, you've seen the output of a recommendation engine. Companies use advanced analytics to make recommendations to potential customers based on the current product that a customer is looking at (using past customers' purchasing history). Exhibit 11.6 shows an example of product recommendations made by **Walmart.com** when a customer looked at a stuffed dog toy. You'll notice that the product recommendations all share certain qualities with the original searched-for item. All of them are dog toys, and all of them are plush or soft. Many of them are animals (sheep, owl, rabbit, duck). In addition, all of the recommendations are in generally the same price range, from $4.97 to $8.53.

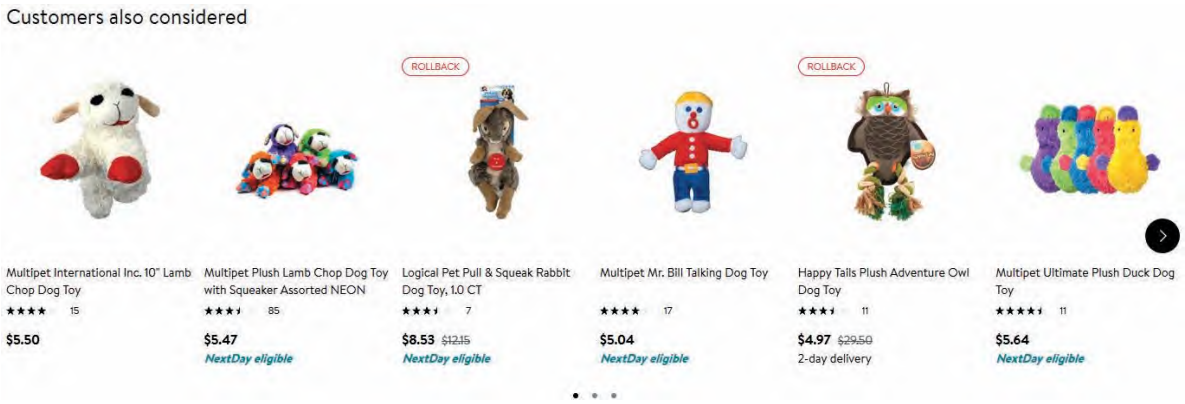


Exhibit 11.6
Walmart.com Product Recommendations

Many consumer-products websites make product recommendations based on the product that the consumer has searched for or looked at. This exhibit shows product recommendations made by Walmart.com after a consumer searched for a specific dog toy. What do all of these recommended products have in common?

Source: Walmart.com.

Recommendation engines based on browsing history are only the tip of the iceberg when it comes to advanced marketing analytics. AI-based recommendation engines can adapt to customer likes and dislikes, immediately react to changes, and make real-time suggestions. For example, the recommendation engine at **Netflix** recommends content based on what you've watched in the past. The recommendations automatically change as your viewing habits change.

[21]Source: Mike Kaput, September 20, 2021, "AI for Predictive Analytics Marketing: What You Need to Know." Available online at https://www.marketingaiinstitute.com/blog/ai-for-predictive-analytics. Accessed 11/19/2021.

AI can also be used for **dynamic pricing**, a pricing strategy used by companies to set flexible prices for products/services based on current market demand, customer purchasing patterns, and competitor prices. AI analyzes a customer's data pattern, determines how receptive that customer is to special offers, and answers the question "What is the customer likely to pay for a product or service?" Based on this information, a company can target an individual customer with the exact level of discount needed to close the sale. For example, **Airbnb** has a well-developed dynamic pricing system that helps property owners determine the appropriate rental price for their property. The system considers many factors, including location, property features, local events, reviews, and photographs.[22]

AI can also help merge data from disparate systems to identify patterns. After blending all relevant data, AI can automatically generate multiple types of analyses, including the sales impacts of specific campaigns, the flow of orders through a website, and the ability of different factors to predict product repurchases. AI can also analyze competitor data to determine what the competition is doing. In short, AI utilizes a broad set of data to create a total view of the customer, resulting in more accurate predictions and prescriptive recommendations.

With this total view of customers, AI can quickly divide customers into **cohorts**, which are groups of customers who share a common characteristic over time. Unlike customer segmentation, which divides customers based on simple characteristics such as age and purchase frequency, cohort analysis more precisely divides customers into smaller groups that share more-specific characteristics over time, such as how long they have been a customer. Marketers can tailor marketing efforts for each cohort. In addition, AI can analyze customer turnover data, find the correlation between the turnover/churn rate and cohort membership, and then recommend cohort-appropriate engagement strategies to prevent customer churn.

turnover/churn rate: The rate at which customers stop buying from a company.

Exhibit 11.7 presents an example of a cohort analysis. In this example, cohorts are defined as users who purchased products initially in the same month. The rates are the repurchase rates in the rest of the months of the year. For example, for the February cohort, 8.25 percent purchased items in December, which is 10 months after the initial purchase. For the November cohort, 5.26 percent purchased items in December.

Count of CustomerID	Column Labels											
Row Labels	0	1	2	3	4	5	6	7	8	9	10	11 Grand Total
Jan	100%	15.57%	17.65%	14.53%	17.30%	16.61%	14.19%	14.88%	15.57%	14.88%	14.19%	13.49%
Feb	100%	15.98%	15.98%	15.98%	9.79%	14.43%	15.46%	13.40%	18.56%	12.37%	8.25%	
Mar	100%	16.91%	19.32%	14.49%	17.87%	12.08%	15.46%	20.29%	14.98%	7.73%		
Apr	100%	11.76%	12.94%	14.71%	12.35%	17.65%	10.59%	8.82%	7.65%			
May	100%	15.25%	15.25%	13.56%	11.02%	21.19%	12.71%	7.63%				
Jun	100%	12.84%	13.76%	16.51%	13.76%	12.84%	11.01%					
Jul	100%	16.47%	17.65%	16.47%	10.59%	7.06%						
Aug	100%	14.86%	18.92%	21.62%	14.86%							
Sep	100%	9.52%	14.29%	15.87%								
Oct	100%	16.28%	6.98%									
Nov	100%	5.26%										
Dec	100%											
Grand Total	100%	14.12%	15.16%	14.05%	12.17%	12.24%	10.29%	9.39%	8.69%	5.77%	3.96%	2.71%

Exhibit 11.7
Cohort Analysis: Repurchase Rates

In the future, AI will help companies dynamically segment the market in increasingly granular segments of similar customers to build fully customized products and to anticipate demand for products based on market signals. For example, in the fashion industry, AI could design new clothes, post them on a website, manufacture the purchased clothes in small batches, and then alter the design based on customer feedback.[23] Ultimately, each customer will receive individually tailored marketing and products.

[22]Source: Rebecca Sentence, 2019, "15 Examples of Artificial Intelligence in Marketing." Available online at https://econsultancy.com/15-examples-of-artificial-intelligence-in-marketing/. Accessed 11/15/2021.
[23]Source: Rebecca Sentence, 2019, "15 Examples of Artificial Intelligence in Marketing." Available online at https://econsultancy.com/15-examples-of-artificial-intelligence-in-marketing/. Accessed 11/15/2021.

ACCOUNTING ANALYTICS

From an accounting perspective, advanced technologies involve automating tasks that were previously labor intensive. For example, accountants used to spend hours reading contracts and agreements. Today, AI systems with natural language processing capabilities, such as **IBM**'s Watson Analytics, can interpret contracts/agreements, identifying key terms and performing risk assessment and other tasks. **Risk assessment** is the process of analyzing events, transactions, or other items that may negatively impact the company.

AI can help the tax function monitor compliance with tax laws and help structure transactions/contracts to reduce future tax liability. For example, compliance with multiple tax jurisdictions can be tricky, especially if a company sells in several or many states or countries. AI helps ensure that sales tax is properly recorded and collected, and it can estimate future taxes based on changing economic conditions and even weather. AI may also help in determining how company decisions may impact the probability of a tax audit.[24]

MINI CASE: Moving Toward a Digital Audit

The goal of auditing is to ensure that a company records its transactions properly, so that the financial statements are accurate. Historically, auditors inspected documents to answer the question "Which transactions are not recorded accurately?" For example, if the auditor wanted to make sure that the company paid a vendor/supplier $1,000 for supplies, the auditor would ask to see the $1,000 check made out to the vendor. The auditor would then verify that the company recorded that transaction properly.

Because of the work involved, the auditor typically selected a sample, or subset, of the documents or transactions to examine, hoping that the sample was representative of all the transactions. This traditional method is labor intensive, time consuming, and expensive even when the transactions are stored in an ERP/ES system and not on paper. In addition, the procedures were not very effective at detecting fraud inside the company. External auditors detect fraud only 4 percent of the time![25]

AI is changing all of this. For example, **MindBridge** Ai Auditor uses AI to analyze all of the transactions in the audit in minutes to answer the question "Which accounts have the highest risk of not being recorded accurately?" These recommendations allow auditors to focus on high-risk transactions to improve audit efficiency and effectiveness. For example, Exhibit 11.8 shows the top of the Risk Overview dashboard created by MindBridge after Ai Auditor analyzed all of the transactions stored in a company's general ledger. The analysis created three risk buckets. It shows that 2 transactions are considered high risk and are worth $124,000; 57 transactions are considered medium risk and are worth $1.57 million; and 434 transactions are considered low risk and are worth $2.68 million.

The Risk Overview dashboard also creates a color-coded bar chart of the transaction by risk level and month. Exhibit 11.9 shows the number of transactions in November,

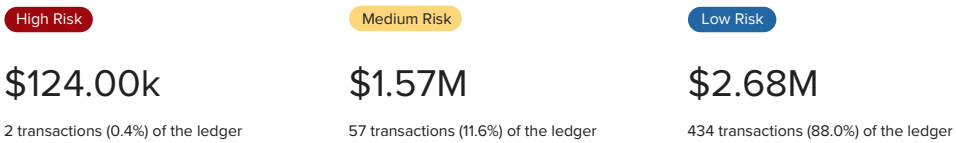


Exhibit 11.8
MindBridge Ai Auditor Risk Overview: Risk Buckets

Source: MindBridge Ai.

[24]Source: "Five Ways Artificial Intelligence Is Changing Tax." May 20, 2017. Available online at http://usblogs.pwc.com/emerging-technology/5-ways-ai-changes-tax/. Accessed 8/29/2019.
[25]Source: ACFE. 2020. *Report to the Nations.* Available online at https://www.acfe.com/report-to-the-nations/2020/ (page 19). Accessed 11/17/2021.

December, and January using the color coding of Exhibit 11.8. For example, the red in December indicates the high-risk transactions.

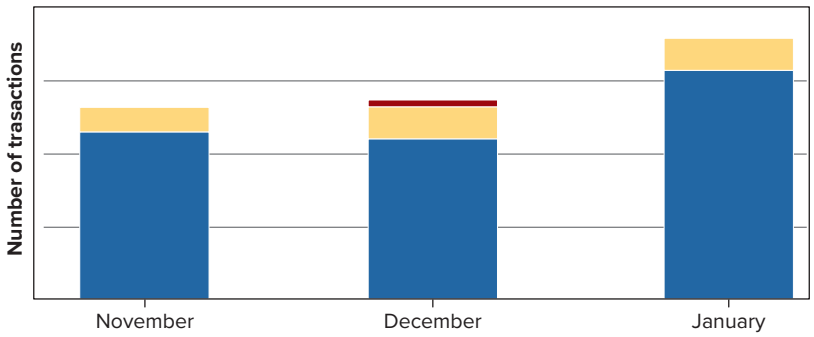


Exhibit 11.9
MindBridge Ai Auditor Risk Overview: Monthly Bar Charts

Source: MindBridge Ai.

The Risk Overview dashboard also automatically creates a circular graph showing the risk by account. Large account categories are in the center, including Assets, Liabilities, Revenues, and Expenses. Specific accounts appear in the outer rings. For example, Exhibit 11.10 shows that the Inventory of raw materials account (account number 14051) has a balance of $241,614.45 with an average risk score of 26.2%, which is considered low risk, as indicated by the light green color.

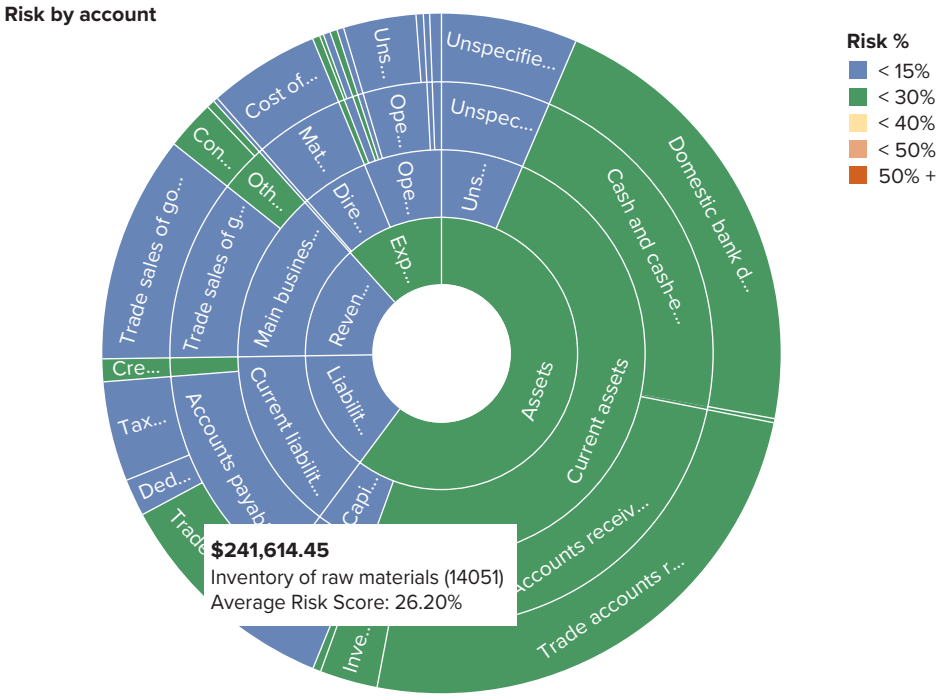


Exhibit 11.10
MindBridge Ai Auditor Risk Overview: Risk by Account

Source: MindBridge Ai.

Finally, the Risk Overview dashboard automatically creates a color-coded scatterplot with the monetary value of the transactions on the X-axis and the MindBridge Risk

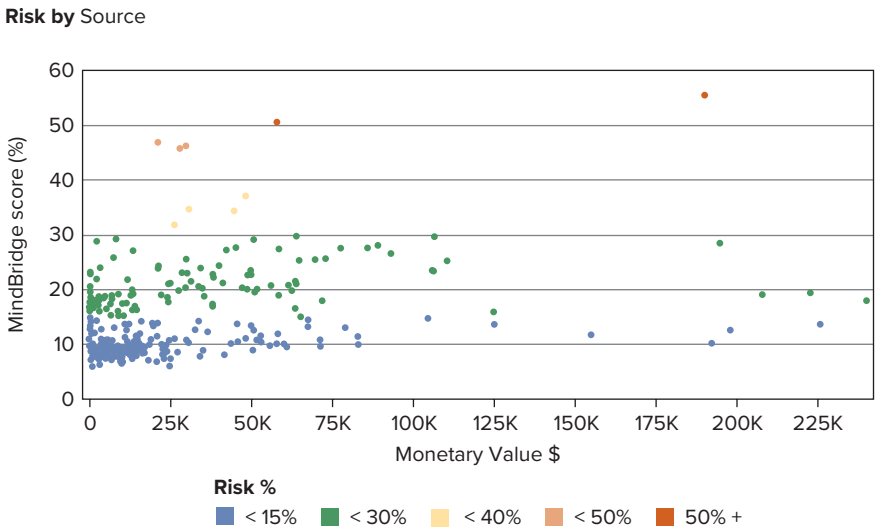


Exhibit 11.11
MindBridge Ai Auditor Risk Overview: Scatterplot

Source: MindBridge Ai.

legend at the bottom. Greater than 50 percent is considered high risk, 30 to 50 percent is medium risk, and less than 30 percent is low risk.

How does the software generate these risk assessments? As Exhibit 11.12 shows, MindBridge Ai Auditor combines the domain expertise of accountants with statistical methods using machine learning and ensemble AI, which improves machine learning results by looking at multiple models.

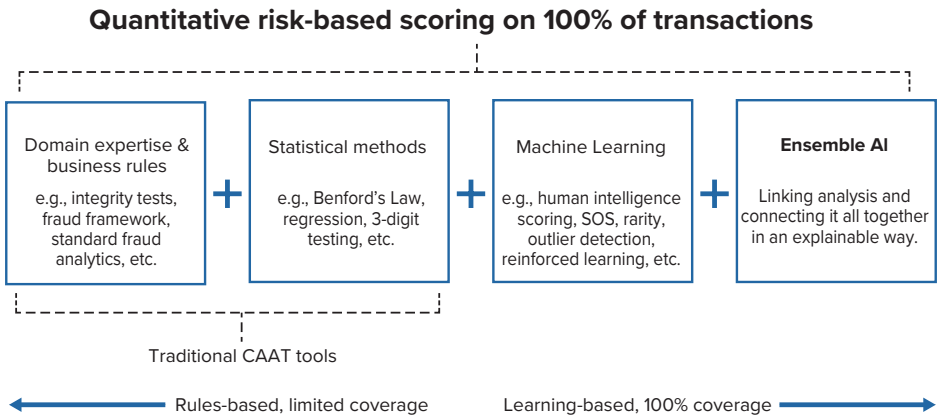


Exhibit 11.12
MindBridge Ai Auditor Quantitative Risk-Based Scoring

Source: MindBridge University Alliance Program slides

Decision-Making with Data

How can an auditor use Exhibits 11.8–11.11 to answer the following questions: Which transactions should be selected for analysis? Which accounts should be investigated? Which month looks the most suspicious? What additional information might an auditor want to see?

AI can also be used to streamline and perform accounting processes. For example, a large accounting firm, **Crowe Horwath**, developed a system to automatically identify complicated insurance claims that historically caused reimbursement problems. It estimates that this system will save clients hundreds of work hours.[26] Another example is **botkeeper.com**, which automatically pays bills, processes payroll, generates invoices, records transactions, and performs reconciliations. The company describes it as "human-assisted automated accounting" and claims that it saves its customers $6 million in bookkeeping costs each year. Exhibit 11.13 presents an overview of the company's processes.

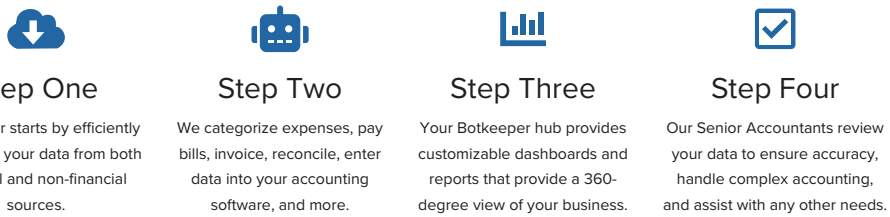


Exhibit 11.13
botkeeper.com's Human-Assisted Automated Accounting

Source: botkeeper.com. Accessed 9/7/2019.

FINANCIAL ANALYTICS

In financial analytics, the three areas that have the biggest potential to leverage the power of AI are personalized financial planning, process automation, and fraud and anti-money laundering detection. Just as there are robocars and robotrucks, there are now robo-advisers. **Robo-advisers** suggest customized investment solutions for everyday consumers by dynamically building and optimizing investments to match investor goals, mimicking the customized investment advice that high-net-worth clients get from their broker. AI can also automate the process of buying/selling stocks and analyzing company information, using some of the techniques outlined in Chapter 9 in combination with other Big Data.[27]

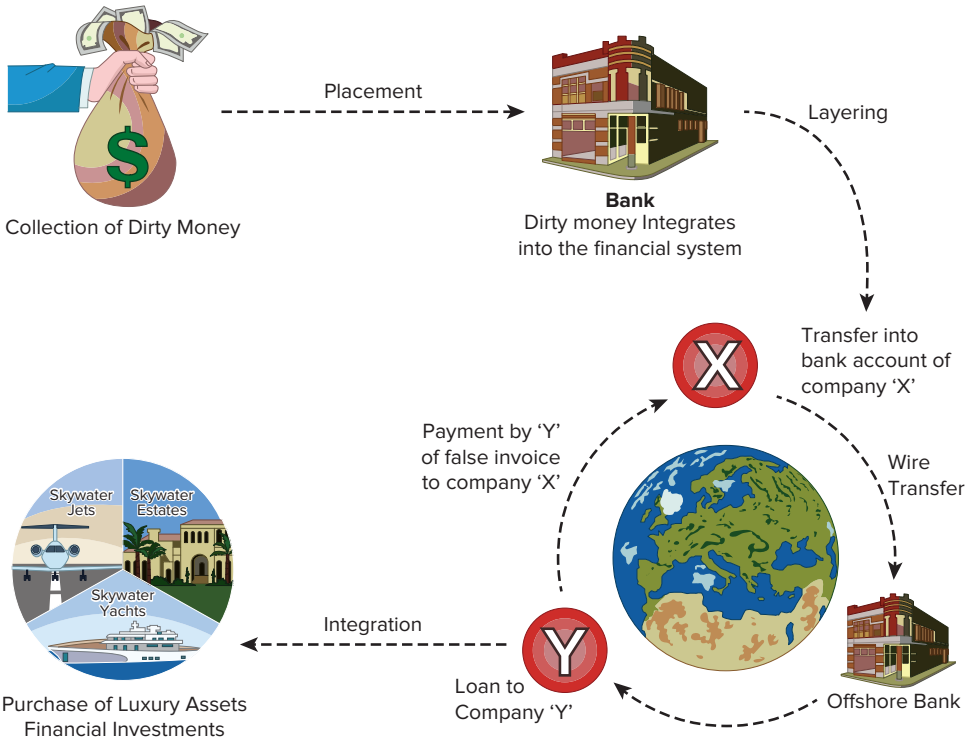
Money laundering is the illegal process of taking "dirty" money generated by criminal activities and making it look "clean" by passing it through a series of business transactions and bank transfers. The goal is to hide the real source of the money from the legal authorities. After the money appears "clean," the criminals withdraw the money from a legitimate bank account. Exhibit 11.14 summarizes a typical money-laundering scheme.

Traditionally, banks used a great deal of resources to detect money laundering, and the majority of the identified potential cases ended up not being cases of money laundering. AI allows banks to better identify money laundering by analyzing internal bank data in conjunction with publicly available data. For example, AI can detect suspicious patterns (not visible to humans) by building a customer's network of acquaintances based on millions of data points from a wide variety of sources, including social media,

[26]Source: Sarah Ovaska-Few, 2017, "How Artificial Intelligence Is Changing Accounting." Available online at https://www.journalofaccountancy.com/newsletters/2017/oct/artificial-intelligence-changing-accounting.html. Accessed 8/26/2019.
[27]Source: PwC, "Smart Money: AI Transitions from Fad to Future of Institutional Investing." Available online at http://www.wavelengthcapital.com/documents/FG/wave/library/591022_pwc-fsi-whitepaper-artificial-intelligence-investing.pdf. Accessed 8/15/2022.

Exhibit 11.14
Money-Laundering Scheme Overview

Source: Fraudswatch, October 19, 2016, "Money Laundering: Corey Earl Engelen Indictment to Money Laundering." Available online at https://www.fraudswatch.com/money-laundering-corey-earl-engelen-indictment-to-money-laundering/. Accessed 8/29/2019.



tax returns, real estate holdings, and Wi-Fi internet protocol addresses accessed by the customer.[28] Transactions with a high-risk person or organization may trigger a money-laundering investigation.

MINI CASE: Social Network Analysis to Detect Money Laundering

For over five years, Sang Bum "Ed" Noh orchestrated schemes to avoid import tariffs and launder money through a company named **Ambiance Apparel**. Ambiance Apparel is actually the operating name Noh used for two separate companies, Ambiance U.S.A., Inc. and Apparel Line U.S.A., Inc. The first scheme had foreign manufacturers prepare a real sales invoice and a lower fake sales invoice for imported goods. The lower fake sales invoice was used to fraudulently reduce import taxes to the tune of $17.1 million. In addition, Ambiance (1) did not document cash transactions of more than $10,000 and (2) faked sales payment documents that excluded cash sales to avoid income taxes. In 2020, Noh pleaded guilty to tax fraud, money laundering, and customs violations. He was fined $118 million and faces prison time.

However, Ed's crimes may not stop there. Exhibit 11.15 answers the question "What is Ed Noh's business network?" All of the companies in the network graph are registered at the same Los Angeles address. "In Y Noh" is the registered agent for the companies.[29] A registered agent is the person whom the company designates to receive process notices, government correspondence, and compliance-related documents.

[28]Source: Lisa Quest, Anthony Charrie, Lucas Jongh du Croo de Jongh, and Subase Roy, August 9, 2018, "The Risks and Benefits of Using AI to Detect Crime." *Harvard Business Review* (August 9). Available online at https://hbr.org/2018/08/the-risks-and-benefits-of-using-ai-to-detect-crime. Accessed 11/18/2021.
[29]Source: Stacy Hubert, September 15, 2020, "Analysis Uncovers Additional Companies in Fraud and Money Laundering Network Not Included in DOJ Indictment." *Sayari.* Available online at https://sayari.com/resources/analysis-uncovers-additional-companies-in-fraud-money-laundering-network-not-included-in-doj-indictment/. Accessed 8/15/2022.

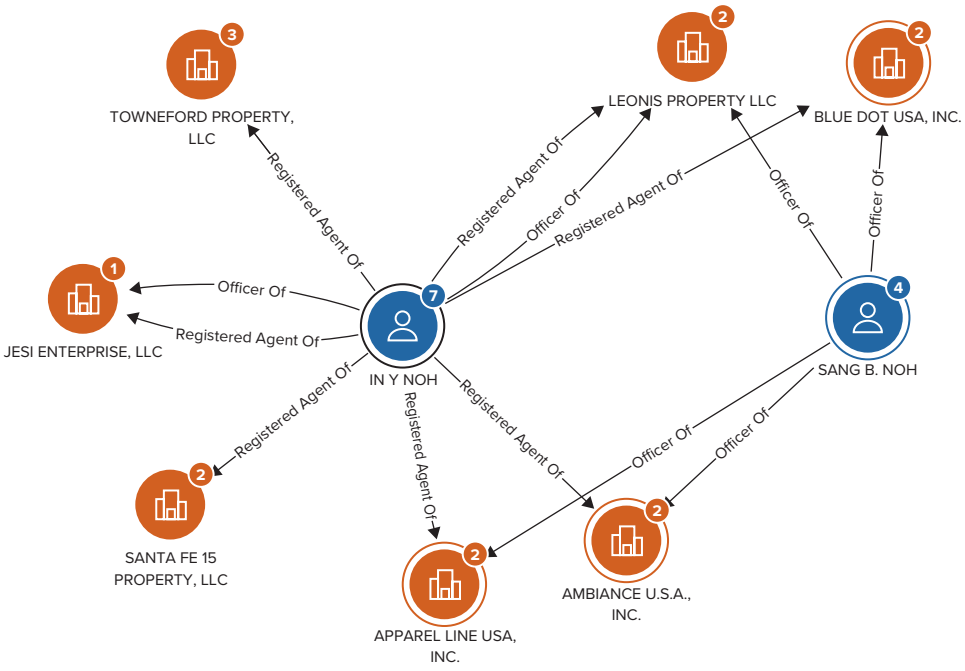


Exhibit 11.15
Money-Laundering Social Network Graph

Source: Stacy Hubert, September 15, 2020, "Analysis Uncovers Additional Companies in Fraud and Money Laundering Network Not Included in DOJ Indictment." *Sayari.* Available online at https://sayari.com/resources/analysis-uncovers-additional-companies-in-fraud-money-laundering-network-not-included-in-doj-indictment/. Accessed 11/18/2021.

Decision-Making with Data

Based on the social network shown in Exhibit 11.15, do you think the money-laundering investigation should be extended beyond just Apparel Line U.S.A., Inc. and Ambiance U.S.A., Inc.? How likely is the possibility of additional fraud? What would you suggest that fraud examiners investigate?

The applications of AI in the financial services industry do not end there. ANNs are also being used to predict bankruptcy, to evaluate credit, and to make loan decisions. All of these applications are part of the **fintech (financial technology)** revolution, in which technology is used in personal and commercial financial applications. Examples of fintech include not only robo-advisers, back-office applications, and commercial applications like the ones described above, but also the use of Bitcoins (a type of digital currency, discussed in more detail later in this chapter) and the use of smartphones to do banking.[30]

OPERATIONS ANALYTICS

To help with their operations, companies can use advanced technologies in many different ways. For example, advanced technologies are automating tasks that were previously labor intensive. Companies also use advanced technologies to help quantify uncertainty and risk, to assist with planning and strategy, and to reimagine the entire business.

[30]Source: J. Kagan, 2019, "Financial Technology—Fintech." *Investopedia.* Available online at https://www.investopedia.com/terms/f/fintech.asp. Accessed 8/29/2019.

HR ANALYTICS[31]

From a human resource (HR) perspective, AI can help a company recruit and hire good employees by delivering personalized messages to prospective new hires and by contacting prior applicants for new postings that better fit their qualifications. This is a huge advancement; historically, former applicants were not considered unless they applied again for a different job.

An AI **chatbot** (short for **chatterbot**) is software that simulates the conversation of a real person, and it can assist HR managers. For example, a chatbot may be able to interview prospective candidates, answer questions about the job post, and collect feedback from and information about the candidates**.** Some chatbots are rules-based and can answer basic questions and respond to programmed commands, while other chatbots are AI-based, understand language, and respond appropriately. An example of a rules-based chatbot is the automated system you interact with on the phone when you call a business and are asked to press buttons or respond verbally to questions. For your balance, please say or press 1 . . .

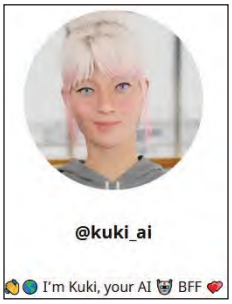


Exhibit 11.16
Meet Kuki

ICONIQ AI

Other chatbots use **natural language processing (NLP)**, a branch of AI that helps computers understand, interpret, and generate human language. If you'd like to experience an AI-based chatbot, you can interact with Kuki, the world's best conversational chatbot. Not only can this chatbot talk, she can also play games and do magic tricks. Exhibit 11.16 presents Kuki.

Once a company hires an employee, AI (including chatbots) can be used to answer the employee's questions about company resources and policies. AI can also act as a company mentor, providing training and support as necessary; schedule meetings by analyzing employees' availability from multiple calendars; and help handle HR compliance cases.[32]

To determine whether a chatbot is helping a company achieve its objectives, companies can use chatbot analytics to measure the bot's performance. For example, companies can assess employees' perceptions and satisfaction with a chatbot as well as track the number and types of issues resolved. Companies can also assess the outcomes of the chatbot's mentoring, training, and support by analyzing employees' ultimate actions. For example, Exhibit 11.17 presents a hypothetical 3D stacked column graph of HR chatbot usage answering the questions "How many employees used the HR chatbot during the first ten days of August?" and "How many employees had a positive interaction with the HR chatbot?" Exhibit 11.17 reveals that daily usage varies from a few interactions to almost 700 interactions, with most employees having a positive opinion of the interaction.



DATA VISUALIZATION

HR managers may want to answer the question "Which words do employees use the most frequently when they chat with the human resources bot?" One visualization technique for presenting this text analysis is a *word cloud,* which shows the most frequently used words. The larger the word in the visualization, the more frequently the word is used in the analyzed text. For example, in Exhibit 11.18 the two most used words are "contribution" and "401K."

Another predictive tool that can help in HR management is **survival analysis**, an advanced predictive analytics technique that estimates the probability of an event occurring over time. Survival analysis uses a cumulative survival curve to visualize the estimated probability of an event. The word *cumulative* reflects that reaching a particular point in time (or milestone) requires reaching the prior milestone. Human resources can use survival analysis to estimate employee turnover at different points in time and to develop a hiring strategy to help guarantee employee coverage. Advances in machine

[31]Unless otherwise noted, the source for this section is https://workology.com/artificial-intelligence-recruiting-human-resources/.

[32]Note: Chatbots can also be used in areas other than human resources. For example, they can be used to interact with customers to improve sales and support. Chatbots and chatbot analytics work the same regardless of the area in which the company uses them.

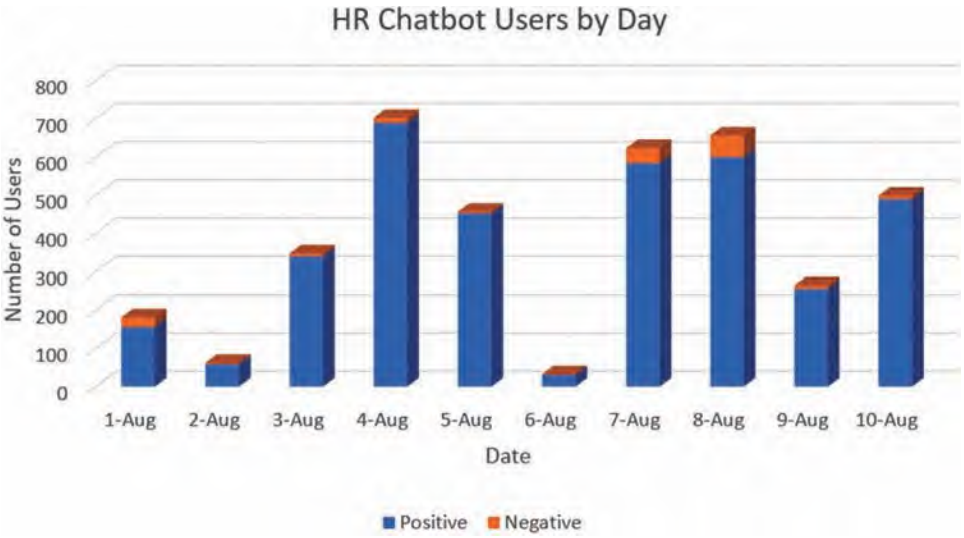


Exhibit 11.17
3D Stacked Column Graph of HR Chatbot Users by Day



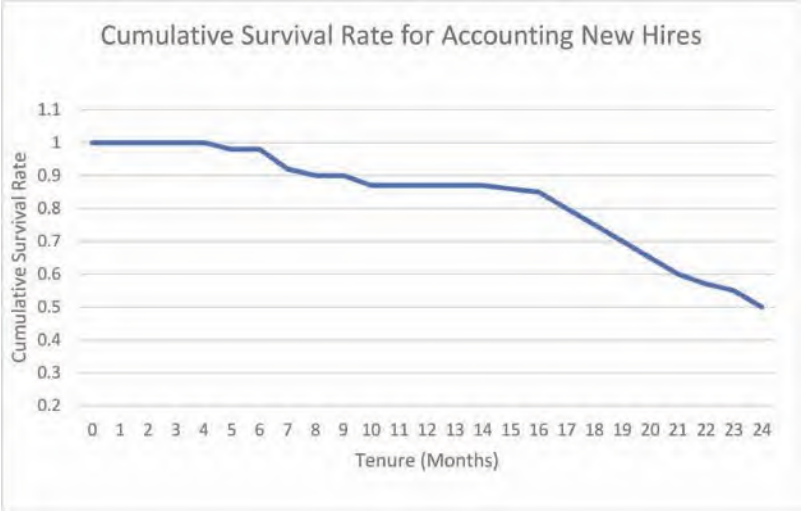
Exhibit 11.18
HR Word Cloud for Employee Interactions with the HR Chatbot

learning may help improve the accuracy of current survival models and analyze drivers of turnover so that management can try to reduce turnover or even hire employees with a lower risk of turnover.

Exhibit 11.19 presents a hypothetical survival line graph for new hires at an accounting firm that answers the question "How many employees will still be employed by the firm two years from now?" The survival graph reveals that at the one-year point (12 months, as measured on the Y-axis), 87 percent of the new hires are still with the firm. At the two-year point, half of the new hires have left. HR managers can use this graph to continuously replenish their staff of qualified accountants. An even better strategy might be to use AI to determine why employees are leaving and identify strategies to minimize turnover, given estimates that it costs at least one-half to two times an employee's annual salary to find a replacement.[33]

[33]Source: Shane McFeely and Ben Wigert, March 13, 2019, "This Fixable Problem Costs U.S. Businesses $1 Trillion." *Workplace.* Available online at https://www.gallup.com/workplace/247391/fixable-problem-costs-businesses-trillion.aspx#:~:text=The%20cost%20of%20replacing%20an,to%20%242.6%20million%20per%20year. Accessed 11/20/2021.

Exhibit 11.19
Employee Survival Analysis for Accounting New Hires



IT OPERATIONS ANALYTICS[34]

We now turn to IT operations analytics. In 2017, **Gartner** coined a new term: **AIOps**, which stands for **Artificial Intelligence for IT Operations.** AIOps uses AI to continuously and automatically collect data, conduct analysis, and report in real time about the *IT infrastructure,* which is all of the hardware, software, and networks used by the company. Given employees' reliance on technology today, a tremendous amount of data is available! AIOps prioritizes problems, highlights infrastructure risks, and identifies the root causes of problems. AIOps helps companies

- monitor system performance and quickly identify exceptions.
- understand what is driving costs and how those drivers change during the day based on system use and hardware capabilities.
- train new employees, freeing up more senior members of IT operations to focus on key decisions.
- minimize application slowdowns and disruptions even at the busiest times of the day.
- eliminate last-minute crises by using alerts to identify potential issues ahead of time.

Exhibit 11.20 presents Gartner's view of AIOps, which relies on Big Data and machine learning. It requires analysis, monitoring, IT service management (ITSM), and automation of actions. Replacing traditional manual processes, AIOps should make systems faster and more available for employees and customers, leading to higher productivity and profitability.

AI may also be coupled with **Monte Carlo simulation**, which uses random sampling in conjunction with statistics to estimate mathematical functions to estimate the probabilities of uncertain events. It is like flipping a coin million times to see what could happen.[35] In AIOps, Monte Carlo simulation can be used to answer the question "How long will the system development project take to complete?"[36]

[34]Source: Morgan Oats, 2018, "What Is AIOps? The Benefits Explained." Available online at https://www.intellimagic.com/resources/blog/aiops-benefits-explained/. Accessed 8/26/2019.
[35]Source: Introduction to Monte Carlo Simulation in Excel. Available online at https://support.office.com/en-us/article/Introduction-to-Monte-Carlo-simulation-in-Excel-64c0ba99-752a-4fa8-bbd3-4450d8db16f1. Accessed 9/7/2019. This article was adapted from Microsoft Excel Data Analysis and Business Modeling by Wayne L. Winston.
[36]This example is adapted from RiskAMP, "Project Planning: An Overview and Example of Using Monte Carlo Analysis in Project Planning." Available online at https://www.riskamp.com/monte-carlo-simulation-in-project-planning. Accessed 11/19/2021.

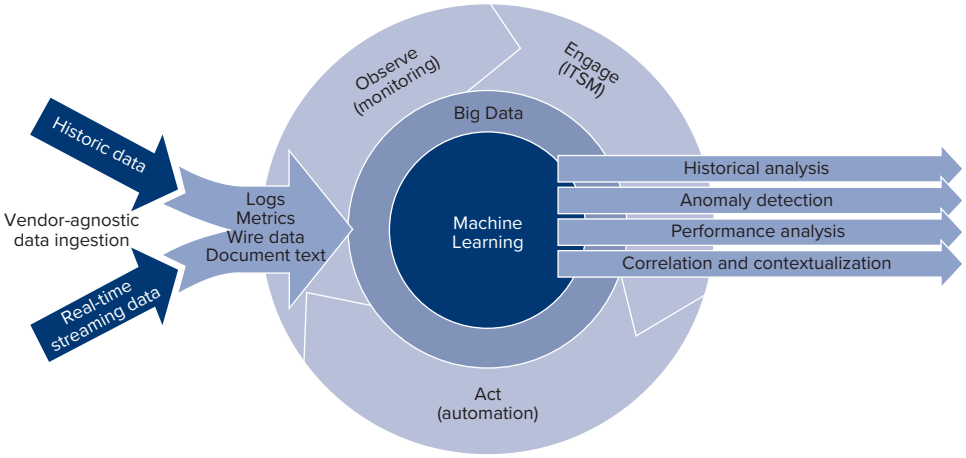


Exhibit 11.20
Gartner: AIOps Platforms

Source: Terry Brown, April 10, 2019, "What Is AIOps? The Benefits Explained." *IT Chronicles.* Available online at https://itchronicles.com/artificial-intelligence/what-is-aiops/. Accessed 11/18/2021.

For each step in the process, estimates for best-case, expected, and worst-case times are made. The *best case* indicates the shortest amount of time to complete the project if everything goes smoothly; *expected* is the expected time to complete the project given typical delays and step times; and *worst case* indicates the longest time to complete the project when every possible delay takes place. Exhibit 11.21 presents a hypothetical example. It shows that the project is expected to take 50 days, but it may take as few as 25 days (best case) or as many as 81 days (worse case). It's unlikely that each step will be completed in the best-case time or the worst-case time, so adding the times together is not the answer. Instead, Monte Carlo simulation runs the analyses based on random estimates of time needed to complete each step. The right-hand column of Exhibit 11.21 presents the time outcome for one trial: 52 days.

Step	Best Case	Expected	Worst Case	One Random Trial Outcome
Step 1	5	10	20	14
Step 2	10	20	30	17
Step 3	7	14	19	14
Step 4	3	6	12	7
Total	25	50	81	52

Exhibit 11.21
Monte Carlo Simulation: IT Project Completion Example

The simulation is run hundreds, thousands, or even millions of times, generating a minimum, average, maximum, and probability of project completion times. If the maximum total time of all of the trials equals 75, then the worst-case scenario is 75 days, not the 81 days shown in Exhibit 11.21, indicating that we overestimated project time for the worst case. We can also generate a graph, like the one in Exhibit 11.22, to show the probability of different completion times. For example, there is a 70 percent probability that the project will be completed in 54 days or less. According to the graph, if you want to be 90 percent certain that the project will be completed when you are scheduling it, then you will schedule the project for 60 days.



DATA VISUALIZATION

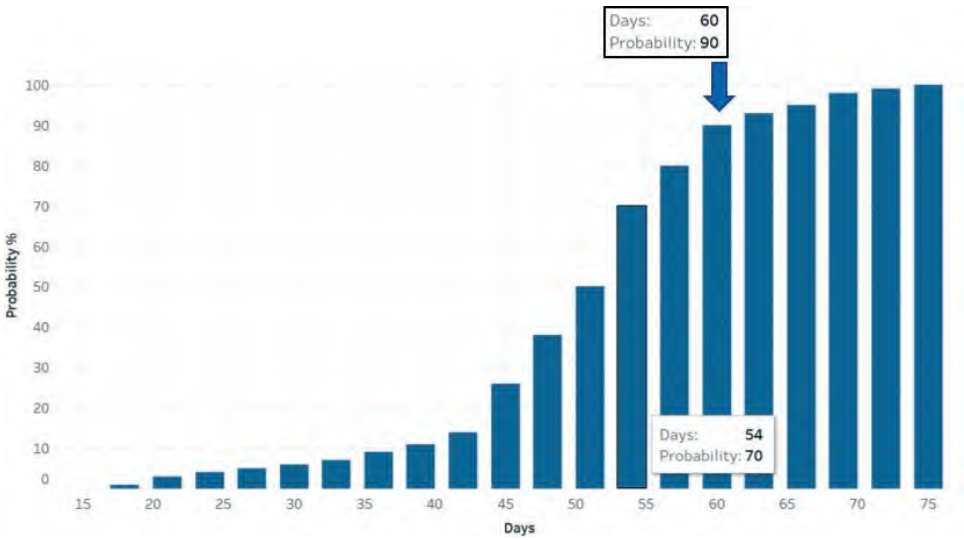


Exhibit 11.22
Monte Carlo Simulation: IT Project Completion Probability Analysis

SOARing to Success

Fun Fact: The History of the Monte Carlo Simulation

Monte Carlo simulation was developed by John Von Neumann and Stanislaw Ulam for the Manhattan Project, which developed the atomic bomb used to end World War II. As part of the Project's research and development, Neumann and Ulam developed the technique to improve decision-making under uncertainty. Basically, the technique allows the user to determine the impact of identified risks on potential outcomes (for example, costs, schedules, and profits). The technique provides you with a range of possible outcomes and the associated probability of each outcome occurring. The technique was named after a tourist district in Monaco, where Ulam's uncle frequently gambled, which seems appropriate as we all know that gambling has a lot of uncertainty!

SUPPLY CHAIN ANALYTICS

"Where drivers are restricted by law from driving more than 11 hours per day without taking an 8-hour break, a driverless truck can drive nearly 24 hours per day. That means the technology would effectively double the output of the U.S. transportation network at 25 percent of the cost."

—RYAN PETERSEN, Techcrunch[37]

Key components of most supply chains are transportation and inventory management. The robocar and robotrucks discussed earlier in this chapter can be combined with completely automated production lines in **lights-out factories** to make and deliver products

[37]Source: Ryan Peterson, 2016, "The Driverless Truck Is Coming and It's Going to Automate Millions of Jobs." Available online at https://techcrunch.com/2016/04/25/the-driverless-truck-is-coming-and-its-going-to-automate-millions-of-jobs/. Accessed 11/18/2021.

to customers with little human assistance (and therefore no need for lighting). For example, since 2001, the Japanese robotics company **FANUC** has been running a lights-out factory that can run for 30 days without human intervention. Similarly, *smart warehouses* are fully automated storage facilities that use robots, automatic guided vehicles, and sensors to move and identify inventory. AI can automate, reinvent, and coordinate functions in a smart warehouse, including unloading inventory, putting inventory on shelves, and picking inventory items from shelves to fill a customer order. For example, **Amazon** uses chaotic or random storage in which non-similar items, such as a garden hose and a John Grisham book, are stored together, with the computer keeping track of each item's position, to optimize inventory storage and retrieval. More recently, *learning warehouses*, which learn and evolve based on AI decisions, are being envisioned to optimize performance.[38]

In addition, advanced analytics can help answer the question "What is the best route for vehicles to deliver products?" For example, **UPS**'s ORION (On-Road Integrated Optimization and Navigation) system routes trucks to minimize left-hand turns, save fuel, and manage the number of trucks required to deliver packages. ORION also modifies the route "on the fly" based on changing weather, road, and accident conditions. In addition, UPS uses AI-driven NPT (network planning tools) to optimize the flow of packages from pickup to delivery. Given that UPS delivers 19 million packages a day using 96,000 trucks, the savings are significant: 10 million gallons of fuel a year and 6 to 8 fewer miles driven per route.[39]

AI can also assist with **demand planning**, which uses market signals to answer the question "What will customer demand be in the next 10–18 months?" Demand planning is important in managing production operations and inventory. **Danone**, a French multinational food company, is using machine learning in its demand planning system, resulting in a 20 percent reduction in forecast error and a 30 percent reduction in lost sales.[40]

In addition to the examples above, advanced analytics is also powering **predictive** (or **conditioned-based**) **maintenance** for equipment and other fixed assets, including robocars and robotrucks. The International Society of Automation estimates that companies lose $647 billion a year from machine downtime. Historically, companies followed planned preventive maintenance determined by the passage of time, hours of usage, or other events. A similar example from our life is the habit of changing a car's oil every 3,000 miles. Unfortunately, these loose guidelines can lead to over- or under-maintaining equipment. A better approach is to ask: When is the optimal time to perform maintenance? AI-based predictive maintenance uses condition-based indicators and alerts to indicate when the fixed asset is at a high risk of failure, but before it fails. For example, the AI system could determine that additional maintenance should be performed when the weather is extremely hot or cold, or the planned route is mountainous.[41] Exhibit 11.23 presents a predictive maintenance dashboard created by IBM.

[38]Source: swisslog, "The Learning Warehouse—The Next Quantum Leap Thanks to Artificial Intelligence." Available online at https://www.swisslog.com/en-us/about-swisslog/our-offering/industry-40/artificial-intelligence-ai-machine-learning-warehouse. Accessed 8/15/2022.

[39]Source: Bernard Marr, June 15, 2018 "The Brilliant Ways UPS Uses Artificial Intelligence, Machine Learning, and Big Data". Adapted from https://www.forbes.com/sites/bernardmarr/2018/06/15/the-brilliant-ways-ups-uses-artificial-intelligence-machine-learning-and-big-data/#6f2ddae85e6d. Accessed 8/26/2019.

[40]Source. Kirsty Donovan, April 5, 2019, "How Supply Chains Are Adopting AI: It Isn't Only for High-Tech Companies and Robots." Available online at https://channels.theinnovationenterprise.com/articles/how-supply-chains-are-adopting-ai. Accessed 8/26/2019.

[41]Source: Uptake (blog), 2018, "How AI is Making Predictive Maintenance a Reality for the Industrial IoT." Available online at https://www.uptake.com/blog/how-ai-is-making-predictive-maintenance-a-reality-for-the-industrial-iot. Accessed 11/18/2021.

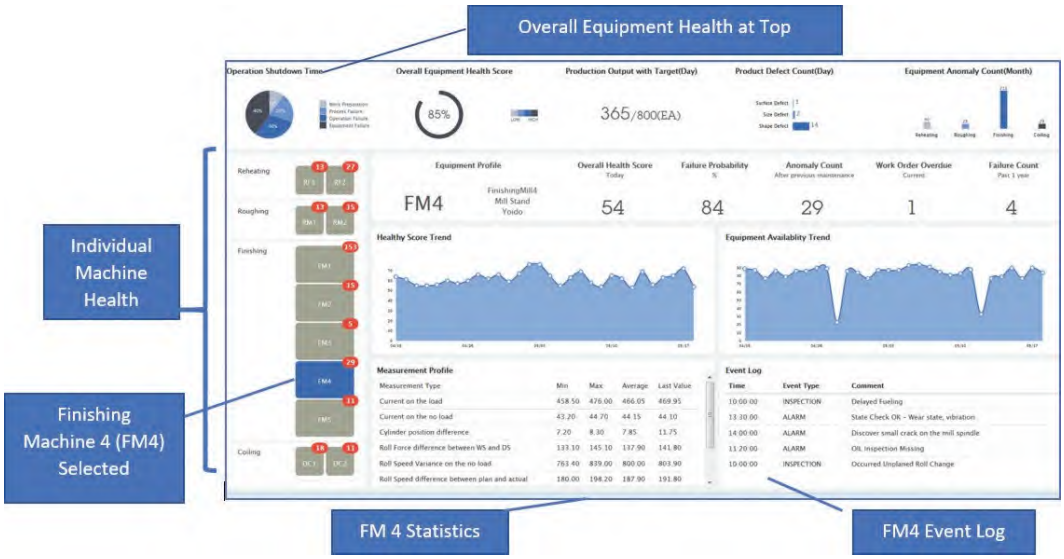


Exhibit 11.23
IBM Predictive Maintenance and Quality Dashboard

Source: Available online at Winn Hardin, May 12, 2016, "Escaping the Reactive Maintenance Rut Through Analytics." *Engineering360.* Available online at https://insights.globalspec.com/article/2646/escaping-the-reactive-maintenance-rut-through-analytics. Accessed 11/19/2021.

LAB CONNECTIONS

LABS 11.1 through **11.4** provide simple examples of methods that can be implemented as part of advanced analytics.

- **LABS 11.1** and **11.2** ask you to work with marketing analytics that can be generated as part of advanced analytics. Lab 11.1 asks you to use Excel pivot tables to analyze cohorts. Lab 11.2 asks you to analyze social media text using a Tableau word cloud.
- **LAB 11.3** asks you to work with a type of HR analytics that can be generated as part of advanced analytics—specifically, Excel's survival analysis to examine the common problem of employee turnover.
- **LAB 11.4** asks you to perform a type of supply chain analytics that can be generated as part of advanced analytics: Monte Carlo simulation in Excel that estimates future demand and contribution margin.

PROGRESS CHECK

5. What do chatbots and robo-advisers have in common? How are they different?
6. It is important to understand the underlying logic of analytics. For example, **UPS** routes its trucks to avoid left turns in order to save time and money. Why do you think avoiding left turns helps increase UPS's performance?

LO 11.4

Describe trending technologies.

11.4 TRENDING TECHNOLOGIES

New technology tools are constantly being developed and implemented to further enhance advanced analytics. This section discusses four trending technologies that companies can use across functions: robotic process automation, cockpits, voice and image analytics, and blockchain analytics.

ROBOTIC PROCESS AUTOMATION (RPA)

Robotic process automation (RPA) is a type of software, similar to that used in robots, that mimics the way people perform different tasks within business processes. It is capable of performing repetitive, routine tasks that do not require insight or understanding—in other words, rule-based tasks. RPA can control desktop applications, collect data, manipulate data, and route data between applications and systems. Once the correct question and the process have been established, RPA can automatically execute the rest of the SOAR model. For example, RPA can be used for automatic **website scraping**, which is the process of extracting data from websites. RPA can also be used for cleaning data, running (predefined) analysis programs at a particular time, and distributing desired output and data visualizations.

The benefits of RPA include fewer errors, shorter processing times, compliance with laws and company regulations, and 24/7 availability. However, as Exhibit 11.24 shows, not all tasks are suited for RPA, which is best used in high-volume, repetitive, stable digital transactions subject to a high error rate when processed by humans. For example, the transactions that are most suited for RPA appear in the HIGH column and have the following characteristics: digital, automatic input; a high volume of processed data; use of a number of software packages; very few errors or exceptions; repetitive process; infrequent process changes; and high rates of human errors in the process.

	SUITABILITY OF RPA	LOW	MEDIUM	HIGH
DATA	Nature of inputs	Physical	Digital Manual entry	Digital Automatic entry
	Volume of processed data	Low	Medium	High
COMPLEXITY	Number of software used by the process	Low	Medium	High
	Frequency of exceptions	High	Medium	Low
STABILTY	Repetitiveness of the process	Low	Medium	High
	Frequency of changes in the process	High	Medium	Low
HUMAN FACTOR	Frequency of human errors in the process	Low	Medium	High

Exhibit 11.24
Suitability of RPA

Source: indeed!, January 24, 2017, "Methodology: How to Handle the RPA Journey?" Available online at https://indeed.headlink-partners.com/2017/01/24/methodology-how-to-handle-the-rpa-journey/. Accessed 11/19/2021.

A current example of RPA is the processing of credit card applications. The RPA ensures that all required documents are received from the applicant; automatically verifies the provided information using credit checks and background checks; and, based on the information provided, determines if the applicant should receive a credit card.[42]

[42]Source: Chandra Gogineni, March 19, 2019, "10 Use Cases of RPA in the Banking Industry." Available online at https://www.comtecinfo.com/rpa/use-cases-of-rpa-in-banking-industry/. Accessed 8/15/2022.

COCKPITS

In previous chapters we've seen how companies use dashboards to present data-based visualizations of business intelligence. Dashboards typically focus on important key performance indicators (KPIs) to help managers understand how the company is performing. They may be interactive, allowing managers to examine the company from different viewpoints.

Cockpits take dashboards to the next level by modeling the underlying business drivers, predicting what is likely to happen under different scenarios, and developing what-if simulations to help managers make corrections and decisions.[43] A cockpit may present and interlink multiple dashboards to provide a clear picture of the entire company and thus allow managers to engage with advanced business analytics to effectively plan strategies, resources, and operations. **Procter and Gamble** has been using cockpits since 2008. The company's cockpit at its Cincinnati headquarters has two 32 × 8 foot screens that immerse decision-makers in the data needed to gain actionable intelligence. The screens help decision-makers focus on exceptions and needed interventions as well as opportunities.[44] Exhibit 11.25 shows a room-sized cockpit.



Exhibit 11.25
Example of a Cockpit

Monty Rakusen/Image Source/Getty Images

VOICE AND IMAGE ANALYTICS

Voice analytics, also known as **speech recognition analytics,** is technology that converts speech to text, analyzes the content, and identifies the speaker's emotion, stress level, and truthfulness based on audio patterns. One type of voice analytics is **voice biometrics**, which creates a unique digital print for a person based on voice characteristics, including

[43]Source: PwC, 2016, "Stepping into the Cockpit: Redefining Finance's role in the Digital Age." Available online at https://www.pwchk.com/en/migration/pdf/finance-role-in-digital-age-jan2016.pdf. Accessed on 8/29/2019.
[44]Source: Brian P. Watson, May 12, 2021, "Data Wrangling: How Procter and Gamble Maximizes Business Analytics." *CIO Insight.* Available online at https://www.cioinsight.com/big-data/data-wrangling-how-procter-and-gamble-maximizes-business-analytics/. Accessed 12/17/2021.

pitch, pace, tone, and vocal cord length, allowing the computer to answer the question "Who is calling?" In the future, voice analytics will likely identify the caller based on voice identification, automatically match callers with agents based on the caller's personality for high levels of service, immediately detect fraud based on speech, and offer suggestions and incentives to a caller based on the caller's history.[45] While voice analytics may lead to better customer service, it does raise several questions, including:



- Is it ethical for companies to identify callers based on their voice?
- Should companies ask their callers for permission to store their voices?
- Is it ethical to profile a customer's personality? Who will develop the criteria to classify a personality?
- Can companies ethically sell their customers' digital voiceprints and personality profiles?
- Should companies take security measures to ensure that digital voiceprints are not hacked?
- What may happen if callers use a voice modulator to hide their real voice?

Image analytics, also known as **computer vision** or **image recognition**, is the process of extracting information from images or videos. It works by AI training the computer via deep learning to "understand" the visual world. Image analytics has a wide variety of applications, including

- Counting inventory on a conveyor belt
- Avoiding an object on the road (for robocars and robotrucks)
- Reading license plates
- Reading bar code tags on inventory
- Identifying brands in photos
- Identifying the faces of employees
- Determining when and where products are used/consumed
- Identifying counterfeit money

Image analytics has been deemed "the future, [as] it's a massive step toward creating truly invisible technology that adapts to users' needs instantly and predicts their future needs with uncanny accuracy."[46] Just think of the applications that can be created when voice and image analytics are seamlessly combined!

BLOCKCHAIN ANALYTICS

A relatively new technology used in various applications is **blockchain**, which creates a distributed ledger based on peer-to-peer networking technology and cryptography that automatically confirms and records transactions, virtually in real time, and is very difficult to change. Let's break down that definition into simple language. Blockchain consists of records or blocks (hence, the block part) in which records/blocks are stored in a continuous linked chain (hence, the chain part). Blocks contain data about something.

[45]Source: Saurabh Goorha and Raghuram Iyengar, June 13, 2019, "Voice Analytics and Artificial Intelligence: Future Directions for a Post-COVID World." Wharton Customer Analytics White Paper. Available online at https://wca.wharton.upenn.edu/white-paper/voice-analytics-and-artificial-intelligence-future-directions-for-a-post-covid-world/. Accessed 11/19/2021.

[46]Source: Hovhannes Avoyan, April 26, 2018, "Technology That Sees the World: Welcome to the Future of Computer Vision." Available online at https://www.forbes.com/sites/forbestechcouncil/2018/04/26/technology-that-sees-the-world-welcome-to-the-future-of-computer-vision/#35007a5f33cb. Accessed 11/19/2021.

For example, a block may record a transaction that took place between a supplier and a company that purchased inventory from the supplier. Each transaction could be a different block in the chain in chronological order. The collection of blocks is sometimes called a *digital ledger.*

To get access to the blockchain, the blockchain members interact via a computer network connection to share information. All of the blockchain members, who are called *peers,* have access to the entire blockchain, and each peer stores its own copy of the blockchain, which makes the ledger both "distributed" and decentralized because one party is not in charge. The blockchain can be public so that everyone can join and participate in the network. Or, if privacy and security are concerns, the blockchain can be private, requiring an invitation or permission to join as defined by the network starter. **Bitcoin,** described below, is an example of a public blockchain, while **MedRec** for electronic health records and **Hyperledger** for blockchain deployments are private blockchains.

The best-known blockchain application is **Bitcoin**, which is a type of electronic cash, digital currency, or *cryptocurrency* (a virtual currency that operates independent of governments and banks). Created in 2007 by a person or group of people using the pseudonym Satoshi Nakamoto, Bitcoin is the oldest and most widely known cryptocurrency as well as the first blockchain database. Other cryptocurrencies are **Ethereum**, **Cardana**, **Binance Coin**, **Tether**, **Sol**, **XRP**, **Dogecoin**, **DOT**, and **USD Coin**. Some cryptocurrencies set the maximum supply of coins; others do not.[47] For example, Bitcoin will have 21 million coins slowly released into the supply until 2136, while Dogecoin and Ethereum do not have a limit.

Cryptocurrencies offer the following benefits:

- enhanced privacy and security.
- low transaction costs.
- quick, permanent transactions.
- lower fraud risk.
- worldwide access.

Cryptocurrencies also have cons. Specifically, cryptocurrencies:

- can make it easier to avoid paying taxes.
- can be difficult to value for taxation purposes.
- are not universally accepted by all vendors.
- require a lot of energy. It is estimated that the energy used by Bitcoin for one transaction is equivalent to 38 house-days' worth of electricity.[48]
- experience heavy market volatility. Since 2009, the value of one Bitcoin has fluctuated from $0.00099 to over $60,000. (To see the current value, go to https://www.buybitcoinworldwide.com/price/.)
- have no system in place to protect their value.
- cannot be recovered if they are lost or stolen.
- are not always completely anonymous.
- are software-based. Thus, cryptocurrency may have software bugs or may be hacked at some point.

[47]Source: "Cryptocurrency Pros and Cons." *CryptoCurrency Facts.* Available online at https://cryptocurrencyfacts.com/cryptocurrency-pros-and-cons/. Accessed 12/17/2021.
[48]Source: Carl Beekhuizen, May 18, 2021, "Ethereum's Energy Usage Will Soon Decrease by 99.95%." Ethereum Foundation Blog. Available online at https://blog.ethereum.org/2021/05/18/country-power-no-more/. Accessed 12/17/2021.

SOARing to Success

Be Careful with That Password . . .

Have you ever forgotten your password to an online account? Typically, you can just reset your password and go on your merry way. However, that is not the case with cryptocurrencies. For example, in 2011 Stefan Thomas, a former Bitcoin trader living in San Francisco, was given 7,002 Bitcoins that may be worth hundreds of millions of dollars now. Passwords protect and provide access to Bitcoins. Stefan wrote his password to his Bitcoins on a piece of paper, which he lost years ago. The account gives him 10 password attempts before the Bitcoins get encrypted, or deleted and inaccessible forever. He has only two password attempts left. Stefan is not the only Bitcoin millionaire who has forgotten a password. So, if you have cryptocurrency and forget your password, you may just have to forget your crypto cash!

Let's now look at how a blockchain works. The first block in the chain is called the *genesis block*. To add additional blocks, blockchain outsiders, called *miners*, solve complex math problems to verify that the block is a valid in exchange for some form of compensation, such as cryptocurrency or information. This mining process is called *proof-of-work*. (Some blockchains use a different approach than mining, but with the same verification purpose.) Once the block is verified, it is added to the blockchain using a cryptographic hash. Cryptography uses complex mathematical codes to store and transmit data between entities with the goal of ensuring that the data can be read only by authorized participants. A *hash* is a function that turns data into a digital fingerprint of data of a standard length. Each block has a unique 256-bit hash that is a combination of letters and numbers that is 64 characters in length. So, if you change any of the data in a block, the hash changes.

In addition to having its own hash, each block contains the hash of the prior block, creating a chain. If information is changed in a block, the hash changes make it not match the original hash, breaking the chain. The use of hashes makes it difficult to change the data because any data changes break the blockchain hashes. This property of blockchains is often called *immutability*.

To make sure members act appropriately, blockchains may have embedded **smart contracts**, which are self-executing legal contracts programmed in the blockchain. Exhibit 11.26 demonstrates how blockchain works. P2P stands for peer-to-peer network.

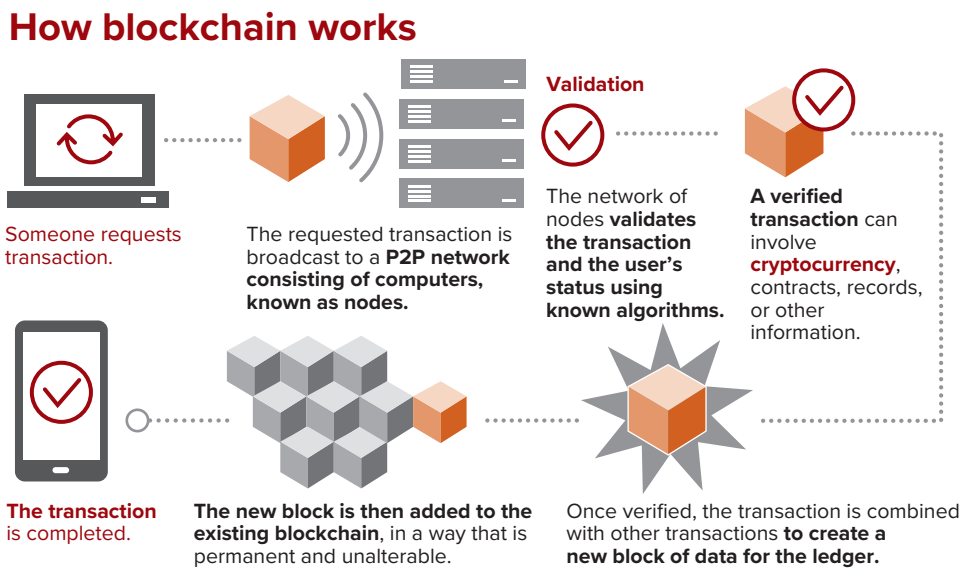


Exhibit 11.26 How Blockchain Works

Source: PwC, 2021, "Bitcoin, Cryptocurrency, Blockchain … So What Does It All Mean?" Available online at https://www.pwc.com/us/en/industries/financial-services/fintech/bitcoin-blockchain-cryptocurrency.html. Accessed 11/19/2021.

A P2P connects computers directly to share data and resources. Each computer is a "peer" with functions that contribute to and support the network. This type of IT infrastructure replaces the old client-server network model where only a centralized computer, the server, provided resources to client computers.

There are two important things to remember about blockchain. First, not all transactions belong on a blockchain. **EY** has identified five questions to ask to determine if a blockchain is the right solution for a business problem. Those who want to implement a blockchain should answer "Yes" to at least three of these following questions:[49]

1. Are multiple parties involved?
2. Is trust between parties an issue?
3. Is it critical to have a detailed, tamper-proof transactional record?
4. Is the group securing the ownership or management of a finite source?
5. Does the network benefit from transparency?

Second, it is important to remember that the mining process does not verify the data inside the block. Rather, it just verifies that the block is valid (agreed upon by all members). So, blockchain content is still potentially:[50]

- inaccurate (includes mistakes and errors).
- unauthorized.
- used for fraudulent or illegal activities, such as money laundering.
- linked to side agreements that are not included in the chain.
- incorrectly reported, summarized, and classified in a company's financial statements.

Even with these two caveats, blockchain has some exciting uses. For example, blockchain can be used to:

- track goods as they go through the supply chain. **IBM**'s Food Trust, which is used by many food companies, records transactions at every step in the supply chain.[51]
- help prevent counterfeiting of luxury goods such as diamonds. **Tracr** tracks a diamond's journey from its source to the ultimate sale.[52]
- store medical records safely. **ScienceSoft** helps healthcare providers build/share blockchains to store and maintain patient records.[53]
- pay for purchases at brick-and-mortar stores and online. **Etsy** and **DISH** as well as some U.K. and U.S. **BMW** dealerships accept Bitcoin.
- raise capital. Companies can sell security token offerings (STOs), which allow investors to purchase digital ownership in a company and/or its assets.
- purchase property. **SMARTRealty** uses custom, agreed-upon smart contracts for property rentals, purchases, and sales.[54]

[49]Source: EY, 2017, "Overview of Blockchain for Energy and Commodity Trading." Available online at https://inperu.pe/boletin/2018/mayo/EY-overview-of-blockchain-for-energy-and-commodity-training.pdf. Accessed 11/21/2021.
[50]Source: K. Tysiac, 2018, "How Blockchain Might Affect Audit and Assurance." *Journal of Accountancy* (March). Available online at https://www.journalofaccountancy.com/news/2018/mar/how-blockchain-might-affect-audit-assurance-201818554.html. Accessed 11/21/2021.
[51]Source: IBM, "IBM Food Trust: A New Era in the World's Food Supply." Available online at https://www.ibm.com/blockchain/solutions/food-trust.
[52]Source: Tracr. Available online at https://www.tracr.com.
[53]Source: ScienceSoft. "Blockchain for Personal Health Records." Available online at https://www.scnsoft.com/healthcare/blockchain-for-personal-health-records.
[54]Source: SMARTRealty. Available online at https://smartrealty.io/.

- buy/sell art. **Maecenas** allows art aficionados to buy and sell a share in 100% authentic artworks by famous artists such as Matisse, Picasso, Rubens, and Warhol.[55]
- pay a variety of insurance claims automatically after a specified event, such as trip cancellation. A now-defunct company called **Fizzy** connected to global air traffic databases and automatically paid customers if the customer's flight was delayed more than two hours.[56]

Blockchains can also impact business analytics.[57] Unlike most data sets, blockchains include a chronological record of every transaction/interaction between entities, allowing deeper (and complete) analysis by **blockchain analytics**. For example, blockchain analytics can identify money laundering, detect fraud, and verify customers by reconstructing hidden patterns in the transactions. Blockchain analytics can also help predict future spending patterns to help speed up business transactions, and it can identify, almost in real time, any changes in blockchain activities. Quickly identifying changes as they are happening allows companies to take immediate corrective actions if necessary.

LAB CONNECTION

LAB 11.5 goes into depth on how blockchain works.

PROGRESS CHECK

7. How is the cockpit in Exhibit 11.25 different from a traditional dashboard?
8. How can blockchain automatically pay an insurance claim for a canceled flight?

11.5 DATA-DRIVEN ORGANIZATIONS AND THE FUTURE OF BUSINESS ANALYTICS

LO 11.5

Describe data-driven organizations and the future of business analytics.

Companies want to serve their stakeholders as well as possible. One way to achieve this goal is to become a data-driven organization. **Data-driven organizations** are companies that use Big Data and business analytics to effectively answer business questions. At these companies, decisions are based on evidence or information gleaned by analyzing huge amounts of data. The impact of becoming a data-driven organization can be huge: Compared to other organizations, data-driven organizations are 23 times more likely to acquire customers, 6 times as likely to retain those customers, and 19 times as likely to be profitable.[58]

A 2021 survey found that only 24 percent of company respondents categorized their organization as data-driven—down from 38 percent in 2020.[59] In other words, most

[55]Source: Maecenas. "The Art Investment Platform." Available online at https://www.maecenas.co/.
[56]Source: Raj Shroff, 2019, "Blockchain in Insurance: Use Cases and Implementation." *The Startup* (August 21). Available online at https://medium.com/swlh/blockchain-in-insurance-use-cases-and-implementations-a42a00ebcd91. 8/9/2022.
[57]The source for this section is: SAS, 2020, "Blockchain: What It Is and Why It Matters." Available online at https://www.sas.com/en_us/insights/analytics/blockchain.html. 8/9/2022.
[58]Source: A study by McKinsey Global Institute as reported by Gaskell, Adi. October 28, 2016. "Becoming a Data Driven Organization." *Forbes.* Available online at https://www.forbes.com/sites/adigaskell/2016/10/28/becoming-a-data-driven-organization/#54983bea4121. Accessed 11/20/2019.
[59]Sources: Randy Bean, January 6, 2020, "New Vantage Partners Releases 2020 Big Data and AI Executive Summary." *Businesswire.* Available online at https://www.businesswire.com/news/home/20200106005280/en/NewVantage-Partners-Releases-2020-Big-Data-AI. Accessed 11/20/2021; Randy Bean, February 5, 2021, "Why Is It So Hard to Become a Data-Driven Company?" *Harvard Business Review.* Available at https://hbr.org/2021/02/why-is-it-so-hard-to-become-a-data-driven-company. Accessed 11/20/2021.

companies are not leveraging the technology and techniques outlined in this book. There are many potential explanations, including total implementation cost, lack of employee talent, entrenched legacy systems, lack of leadership, and slow-to-change corporate cultures. Another major reason is reducing business analytics to an IT problem and solution. Just implementing the latest technology does not guarantee a company's success. Rather, the successful use of business analytics rests on asking relevant business questions related to a company's current and future strategy, and then using the correct statistical techniques and the correct technology tools to answer those questions. In other words, the tools and techniques explained in this book (along with future advances) must be tailored to meet the organization's needs. Only then can a company truly become data-driven.

LIMITATIONS OF BUSINESS ANALYTICS

Unfortunately, business analytics may not always be able to provide the insights desired by a company. This lack of results may have several causes.

First, the entire company may not buy into business analytics, in which case the wrong questions may be asked, and the results may not be communicated to the right decision-makers.

Second, management may believe that business analytics is easy to introduce throughout the company. However, it usually takes time to implement and tweak the business analytics process to provide valuable insights. Business analytics is not a quick fix.

Third, the data needed to answer the questions may not be collected or may be of poor quality (inaccurate, incomplete, incorrect formats, or just proxies). AI systems are trained by data inputs. If only a subset or nonrepresentative sample of data are used are used as inputs, the results of business analytics, including analytics powered by AI, may be misleading or inaccurate. For example, if pictures are used as an input, the angle at which pictures are taken could impact the AI analytics results. Likewise, if inputs are not labeled consistently because of different viewpoints and different attention to detail, the results may be misleading as well.

Fourth, the formulation/phrasing of the question and the evaluation of the model's performance may create misleading results.

Fifth, some data analytics problems cannot be solved by current technology because the data set is too big or because the problem is just too complex. Moreover, the larger the data set and its number of characteristics (fields/features/dimensions), the higher the chance of finding misleading patterns in the data. Companies need to be aware of these issues and work to overcome them as they pursue increased profitability and efficiency through business analytics.

THE FUTURE OF BUSINESS ANALYTICS

This chapter has described the use of adaptive/autonomous analytics across the businesses processes as well as some trending technologies. What does the future hold? While no one knows for sure, here are a few things that we think will happen. First, business analytics will no longer be just a catch phrase or an optional part of business planning. Rather, business analytics will become a normal part of doing business, woven seamlessly into and throughout a company's business strategy and processes. Data-driven decision-making will be the norm, not the exception. All employees will be expected to be comfortable using business analytics at some level because human decision-makers are still necessary in most contexts to interpret output, identify issues, and set up and monitor automated processes.

Second, as statistical techniques and technology tools continue to evolve, business analytics will evolve too, generating more powerful actionable intelligence. For example, a

new quantum computer is being developed to solve problems in a different way than current computers do. Although quantum computers will not replace today's computers, they may be able to solve *selected* business analytics problems faster than today's computers and solve problems that currently cannot be solved. For example, a **Google** quantum computer can solve a specific complex math equation in just 3 minutes and 20 seconds, compared to the estimated 10,000 years solve time of a traditional computer.[60] Several companies are currently exploring how quantum computers can improve critical aspects of their business, including **Delta** with air routes, **Daimler** (think **Mercedes-Benz**) with car batteries and manufacturing defects, and **JP Morgan** with option pricing and portfolio management.

However, new statistical techniques and technology tools also create new risks. For example, quantum computers can already break current encryption technologies, potentially leaving personal data exposed to hackers. Employee data, company financial data, and trade secrets may also be compromised. Experts anticipate quantum computers will also break blockchains based on current technologies, leaving digital cryptocurrencies open to theft. And, unlike paper currency, digital money can be there one second and gone the next second. Ideally, quantum computers will help create a new type of encryption technology that will be unbreakable—at least until the next major statistical technique and/or technology tool advancement.

Whatever happens, this is an exciting time to be involved with business analytics. Employees with an aptitude for business analytics will find greater opportunities than ever before. Best of luck on your journey!

PROGRESS CHECK

9. Describe a data-driven organization.
10. Quantum computing can (potentially) be used to optimize performance. Which business processes discussed in this book might be optimized using quantum computing?

[60]Source: Ernest Hamilton, November 22, 2019, "Can Quantum Computing Break the Blockchain?" *Tech Times*. Available online at https://www.techtimes.com/articles/246148/20191122/can-quantum-computing-break-the-blockchain.htm. Accessed 12/18/2021.

Summary

This chapter described several of the most common types of advanced analytics. We described how adaptive/autonomous analytics builds on the foundation of the other four types of analytics. We also described the underlying primary technology, AI (artificial intelligence). Finally, we provided examples of how companies can use advanced technologies to answer important business questions. When implemented correctly, AI applications should generate actionable intelligence, which should in turn have a positive impact on company performance. In other words, AI-based decision-making can be used to reduce costs, increase production, and ultimately increase profits.

Key Terms

adaptive/autonomous analytics or cognitive analytics (*777*) The newest type of business analytics; it mimics human thought processes and (self-)learning to rapidly gain insight into large amounts of data.

algorithm (*777*) A set of computer instructions to perform a calculation or business process.

artificial intelligence (AI) (*780*) A collective term for computer systems that can sense their environment, think, learn, and take action in response to what they're sensing and their objectives.

Artificial Intelligence for IT Operations (AIOps) (*794*) A part of business operations that uses AI to continuously and automatically collect data, conduct analysis, and report in real time about the IT infrastructure (that is, all the hardware, software, and networks used by the company).

artificial neural networks (ANNs) (*782*) A type of machine learning that uses connected processing nodes resembling the human brain.

Bitcoin (*802*) Electronic cash, digital currency, or cryptocurrency that is based on blockchain technology. Bitcoin is the oldest and best-known type of cryptocurrency.

blockchain (*801*) A distributed ledger based on peer-to-peer networking technology and cryptography that automatically confirms and records data, virtually in real time, and is very difficult to change. Simply put, blockchains contain linked, permanent records of interactions between blockchain participants using techniques to keep the data private and secure.

blockchain analytics (*805*) The process of analyzing blockchain data.

chatbot (chatterbot) (*792*) Software than can hold a conversation. It may be rules-based and respond to a few commands, or it can be AI-based and understand language.

cockpits (*800*) Visualizations that model underlying business drivers, predict what is likely to happen under different scenarios, and develop what-if simulations to help managers make corrections and decisions.[61]

cohort (*785*) A group of customers who share a common characteristic over time.

data-driven organization (*805*) A company that uses Big Data and business analytics to effectively answer business questions.

data lake (*781*) A repository for a large amount of both structured and unstructured data for data analysis and reporting.

data mining (*781*) The process of analyzing large amounts of data to identify patterns, anomalies, and relationships.

data warehouse (*781*) A repository that allows a large amount of structured data to be integrated for data analysis and reporting.

deep learning (*782*) Artificial neural networks that are deep (many layered).

demand planning (*797*) Analytics that uses market signals to predict customer demand for the next 10–18 months.

dynamic pricing (*785*) A pricing strategy used by companies to set flexible prices for products/services based on current market demands, customer purchasing patterns, and competitor prices.

fintech (financial technology) (*791*) Technology used in personal and commercial financial applications.

image analytics (computer vision or image recognition) (*801*) The process of extracting information from images or videos.

lights-out factories (*796*) Fully automated production factories requiring minimal human supervision and therefore no lighting.

machine learning (ML) (*781*) Statistical techniques that allow a computer to "learn" from data without being explicitly programmed.

[61]Source: PwC, 2016, "Stepping into the Cockpit: Redefining Finance's Role in the Digital Age." Available online at https://www.slideshare.net/PWC/stepping-into-the-cockpit-redefining-finances-role-in-the-digital-age. Accessed 8/15/2022.

Monte Carlo simulation (*794*) A technique that uses random sampling in conjunction with statistics to estimate mathematical functions to estimate the probabilities of uncertain events.

natural language processing (NLP) (*792*) A form of AI that helps computers understand, process, and generate human language.

predictive (conditioned-based) maintenance (*797*) The use of historical and contextual data to alert a company to conditioned-based maintenance needs.

recommendation engine (*784*) Data-filtering tool that uses algorithms to predict what a user may or may not like.

risk assessment (*786*) The process of analyzing events, transactions, or other items that may negatively impact the company.

robo-advisers (*789*) Digital platforms that provide automated, algorithm-driven financial planning services based on a customer's current financial situation and future goals.

robotic process automation (RPA) (*799*) A type of software that mimics the way people perform different tasks within business processes. It can perform high-volume, repetitive, routine tasks that do not require insight or understanding.

smart contracts (*803*) Self-executing legal contracts programmed in a blockchain.

survival analysis (*792*) An advanced predictive analytics technique that estimates the probability of an event occurring over time.

voice analytics or speech recognition analytics (*800*) Technology that converts speech to text; analyzes the content; and identifies speaker emotion, stress level, and truthfulness based on audio patterns.

voice biometrics (*800*) Technology that creates a unique digital print for a person based on voice characteristics (such as pitch, pace, tone, and vocal cord length).

website scraping (*799*) The process of extracting data from websites.

ANSWERS TO PROGRESS CHECKS

1. Adaptive means that the programs underlying the analytics change as more data about decisions and outcomes are acquired. Autonomous means that the software can learn by and adapt on its own.
2. Adaptive/autonomous analytics can analyze large volumes of information quickly to determine customers' preferences. Currently, companies use this information to make product recommendations. In the near future, fully customized products will likely be available. In the long term, companies may be able to anticipate future demand for specific products based on market signals.[62] At **MetLife**, adaptive/autonomous analytics could monitor calls in real time, offering suggestions to help the representatives improve their performance as well as solutions to the customer's problem.
3. Artificial intelligence is an advanced analytics technique that includes machine learning and artificial neural networks. In other words, machine learning and artificial neural networks are types of artificial intelligence. Therefore, artificial intelligence is the broadest term.

[62]PwC, 2017, "Sizing the Price: What's the Real Value of AI for Your Business and How Can You Capitalise?" Available online https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html. Accessed on 8/21/2019.

4. **Facebook** could use deep learning to recognize uploaded pictures to determine users' preferences for food, brands, and vacation spots. In addition, Facebook could use deep learning to determine social networks. (Many other correct answers are possible.)
5. Both chatbots and robo-advisers can use AI to interact with people. Chatbots are general purpose and able to understand natural language and carry on conversations about many topics. Robo-advisers are special purpose and provide assistance only for investments.
6. When you make a left turn, you typically have to wait for an opening in traffic to make the left turn (the wait time may be reduced by a dedicated left-turn arrow on the stop light). This waiting causes your car to idle longer than when you make a right turn. So, left turns take longer to make than right turns, and they use more gas. In fact, research has shown that "super streets," which require a right-hand turn and then a U-turn to go left, reduce fuel consumption and travel time by 20 percent.[63]
7. A cockpit combines multiple dashboards about the company to give a better overall view of company performance. The cockpit immerses decision-makers in the data needed for actionable intelligence. The enormous screens help decision-makers focus on exceptions and needed interventions as well as opportunities.
8. Flight cancellation insurance executed by blockchain would contain a programmed smart contract that would identify the flight details (such as carrier, flight number, date, and time of departure). The blockchain would automatically monitor the airline flights, and if the flight is canceled, it would automatically send payment to the customer within minutes.
9. Data-driven organizations are companies that use Big Data and business analytics to effectively answer business questions. In other words, decisions are based on evidence or information gleaned by the analysis of large amounts of data.
10. Quantum computer optimization may be able to optimize inventory management, supply chains, shipping, stock portfolios, risk analysis, fraud detection, and the design of products. (Other correct answers are also possible.)

connect

Multiple-Choice Questions

1. (LO 11.1) Which type of analytics answers the question "What action(s) should we take?"
   a. Adaptive/autonomous
   b. Descriptive
   c. Diagnostic
   d. Predictive
   e. Prescriptive
2. (LO 11.1) The types of business analytics are not related.
   a. False
   b. True

[63]Source: D. Spector, 2012 "Why Left-Hand Turns Are Burning a Hole in Your Wallet." *Business Insider* (March 15). Available online at https://www.businessinsider.com/save-money-gas-avoid-left-turns-2012-3. Accessed 1/16/2021.

3. (LO 11.2) Which type of technology is based on the human brain?
   a. Artificial neural networks
   b. Blockchain
   c. Data mining
   d. Website scraping
4. (LO 11.2) Which of the following items would be considered strong AI?
   a. A robot that can beat the best human chess player in the world.
   b. A robot that can perform a bank reconciliation of millions of transactions.
   c. A robot that helps doctors diagnose a medical condition.
   d. A robot that functions like a human.
5. (LO 11.3) What term means using market signals to predict customer demand for the next 10 to 18 months to manage production operations and inventory?
   a. Data mining
   b. Demand planning
   c. Fintech
   d. Predictive maintenance
6. (LO 11.3) Kuki is an example of:
   a. an AI-based chatbot.
   b. AIOps.
   c. a robo-adviser.
   d. a rules-based chatbot.
7. (LO 11.4) A good candidate for robotic process automation is a task with a:
   a. high number of exceptions.
   b. high volume of processed data.
   c. low number of human errors.
   d. process that changes frequently.
8. (LO 11.4) **Procter and Gamble** utilizes a Business Sphere room, which is a conference table surrounded by large screens that display visualizations of the company's data. The room is used to foster collaboration to quickly solve company problems. Which type of technology would be appropriate for the screens in the Business Sphere room?
   a. Blockchain
   b. Cockpit
   c. Data lake
   d. Robotic process automation
9. (LO 11.5) Bitcoin is an application of:
   a. artificial intelligence.
   b. blockchain.
   c. machine learning.
   d. neural networks.
10. (LO 11.5) Which of the following is NOT true of a smart contract?
    a. It is a legal contract between members of the blockchain.
    b. It is programmed into the blockchain.
    c. It is self-executing based on met conditions.
    d. Robotic process automation (RPA) must be used.

Discussion Questions

1. (LO 11.1) Describe how the first four types of analytics are related to the fifth type, adaptive/autonomous analytics.
2. (LO 11.2, 11.3) AI can develop a personal network for banking customers (specifically, all the entities associated with a customer) to help identify potential money laundering. When conducting a financial statement audit of a company, auditors want to ensure that the financial transactions are valid, recorded accurately, and are between independent, non-related parties. How could auditors use AI to help them audit the transactions?
3. (LO 11.2, 11.3) Companies typically use leases to rent property, buildings, and equipment. A lease is based on a lease agreement, which a legal contract between two parties. The tax law determines how the lease should be treated in the company's tax returns. The problem is that tax laws constantly change, affecting how companies record the leases. How could AI be used to help evaluate how the changes in the tax code impact a company's treatment for its leases in its tax returns?
4. (LO 11.3) How can chatbots be used to automate the supply chain?
5. (LO 11.4) What is ETL? (*Hint:* This term was introduced in a Chapter 2.) How can robotic process automation (RPA) help automate the process?
6. (LO 11.5) How could visual analytics help retailers make product recommendations to customers?
7. (LO 11.4) How is voice analytics different than the natural language processing technology behind **Amazon**'s Alexa and **Apple**'s Siri?
8. (LO 11.5) A company has work sites all over a large city. The company pays employees based on the number of hours worked on the sites. The company also has valuable assets at each work site. How could the company use technology to monitor employees and assets?
9. (LO 11.5) Instead of paper or electronic tickets, some airlines now want to use your face as a boarding pass. At the 20 largest airports in the United States, the U.S. Customs and Border Protection Agency is implementing facial recognition systems for international travel. These systems record not only your face but also other biometric data, such as fingerprints and iris scans. The goal is to track all passengers from curb to gate. **Hertz** is also implementing facial recognition and fingerprint scans, which allow customers to check cars out in 30 seconds.[64] What type of technology covered in this chapter is being used to implement these new systems? What are the benefits of such technology to companies and customers? Are there any concerns for customers?
10. (LO 11.1, 11.2, 11.3, 11.4, 11.5) Given all the advanced technologies presented in this chapter, what skills will employers be looking for in new hires?
11. (LO 11.4) **Amazon** has recently filed 17 patents for biometric technology to include in its **Ring** doorbells. (Amazon purchased Ring in February 2018.) The technology would help Ring cameras identify a "suspicious" person based on odor, voice, gait, fingerprints, retina, face, and other personal characteristics. The patents imply that all Ring doorbell cameras in a neighborhood would be interconnected to share the information and build a profile and composite image of the "suspicious" person. One of the patents, the neighborhood alert mode patent, would allow Ring owners of one household to send photos and videos to other Ring doorbells, so each neighborhood Ring doorbell could record the "suspicious" person and create videos that will be merged to create a storybook of activity for the "suspicious" person.[65] This proposed functionality would be added to Ring's current video recording capabilities and audio processors, which (according to some sources) may capture movement and sound up to 200 feet away from the doorbell. Which technologies discussed in this chapter may be incorporated into Ring doorbells? What benefits does the proposed technology offer? Are there any concerns for customers?

[64]Source: Jason Davis, March 14, 2019, "Biometric Screening at Airports Is Spreading Fast, but Some Fear the Face-Scanning Systems." Available online at https://www.nbcnews.com/mach/science/biometric-screening-airports-spreading-fast-some-fear-face-scanning-systems-ncna982756. Accessed 11/20/2021.
[65]Source: Adam Smith, December 16, 2021, "Amazon Ring Working on Doorbells That Detect People by Their Skin and Smell." *Yahoo! News.* Available online at https://www.yahoo.com/entertainment/amazon-

connect

Brief Exercises

1. (LO 11.2) Match the type of advanced analytics technology with its description.

Advanced Analytics Method	Definition
1. Artificial intelligence	a. A form of AI that helps computers understand, process, and generate human language
2. Machine learning	b. A type of AI that uses connected processing nodes resembling the human brain
3. Data mining	c. Statistical techniques that allow a computer to "learn" from the data without being explicitly programmed
4. Artificial neural networks	d. A collective term for computer systems that can sense their environment, think, learn, and take action in response to what they're sensing and their objectives
5. Natural language processing (NLP)	e. The process of analyzing large amounts of data to identify patterns, anomalies, and relationships

2. (LO 11.3, 11.5) Match the application of each advanced analytics method with its description.

Application of Advanced Analytics	Description
1. Image analytics	a. Creates a unique digital print for each person based on voice characteristics
2. Voice analytics	b. Collects data, conducts analysis, and reports in real time about the IT infrastructure
3. Blockchain analytics	c. Extracts information from images or videos
4. Voice biometrics	d. Uses market signals to predict customer demand for the next 10–18 months
5. Demand planning	e. Analyzes data stored in a distributed ledger
6. Artificial Intelligence for IT Operations (AIOps)	f. Converts speech to text, analyzes the content, and identifies speaker emotion based on audio patterns

3. (LO 11.2, 11.4) Match each tool with the correct description.

Tool	Definition
1. Robotic process automation	a. A central, organized database of company data formatted for analysis to support business intelligence
2. Dashboard	b. Extracting data from websites
3. Cockpit	c. Interactive visual display of KPIs
4. Enterprise system	d. Mimics the way people perform repetitive, routine tasks that do not require insight or understanding
5. Website scraping	e. Visualizations that model underlying business drivers, predict what is likely to happen under different scenarios, and develop what-if simulations to help managers make corrections and decisions

4. (LO 11.3, 11.4, 11.5) For each of the business questions below, identify the appropriate technology to help answer the question from the following list:
   a. image analytics
   b. demand planning
   c. dynamic pricing
   d. voice analytics

Business Question	**Technology** (Image Analytics, Demand Planning, Dynamic Pricing, or Voice Analytics)
1. What is the condition of a field of crops?	
2. How will a hot summer affect customer demand for a specific product (for example, sunscreen)?	
3. How much is a customer willing to pay for satellite TV?	
4. How many inventory items are currently in the warehouse?	
5. When should a customer's call be quickly routed to a manager?	

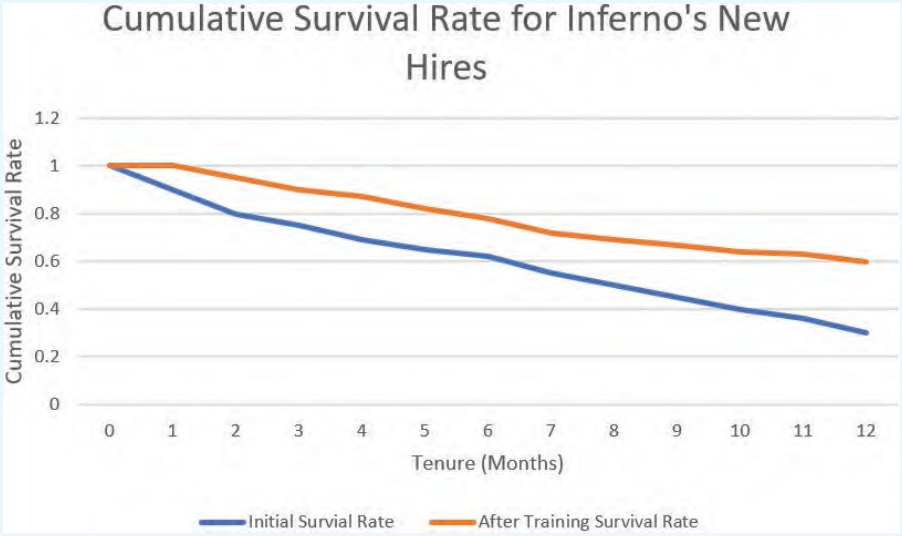
5. (LO 11.5) Match the terms related to blockchain with the correct definition.

Blockchain Term	**Definition**
1. Bitcoin	a. People who solve complex math problems to add another block to a blockchain
2. Distributed	b. Difficult to change without collusion
3. Blocks	c. Programmed contract
4. Cryptography (hashes)	d. An application of blockchain
5. Smart contract	e. Records in the blockchain
6. Immutable	f. Used to link the blocks
7. Miners	g. Data are shared at multiple places

Problems

1. (LO 11.2, 11.3) Labor costs are one of the top concerns of restaurant managers and owners. It takes about 42 days and over $4,000 to hire a new restaurant worker, and most restaurant employees quit due to a lack of training.[66] Let's assume that a local Texas pizzeria, **Inferno's Wood Fired Oven Pizza,** is experiencing high employee turnover and wants to use a data-driven approach to increase employee retention. A year ago, the restaurant introduced a new training program and encouraged managers to help support the staff during busy times. The following data visualization shows retention before and after Inferno's implemented the new training program using a cumulative survival rate curve for employee tenure.

[66]Source: Morgan Harris, November 19, 2021, "Improving Restaurant Employee Retention with a Data-Driven Culture." *QSR.* Available online at https://www.qsrmagazine.com/outside-insights/improving-restaurant-employee-retention-data-driven-culture. Accessed 11/21/2021/.



Answer the following questions about the data visualization.

a. What is a cumulative survival rate curve? Why is it called cumulative?

b. Before the training program, what percentage of the employees were still working at Inferno's two months after starting? Five months after starting?

c. Did the training program improve employee retention? If so, by how much at the 12-month mark after starting?

2. (LO 11.2, 11.3) A small company with a new product collected phrases used in social media about the product. The phrases are presented in the data visualization below.



Answer the following questions about the data visualization.

a. What is the data visualization called?

b. What method of analysis was used to create the data visualization?

c. What is the most frequently used word?

d. Chapter 7 introduced sentiment analysis, which is a technique that determines the emotion behind words. How would you describe the sentiment in the data visualization? What would you recommend the company do in light of the sentiment?

3. (LO 11.2, 11.3) Worldwide shoplifting costs companies $15 billion a year.[67] A Japanese company, **VAAK**, is attempting to stop shoplifting in its tracks. VAAK developed

[67]Source: Arthur Zuckerman, May 30, 2020, "35 Shoplifting Statistics: 2020/2021 Data, Trends & Impact." *CompareCamp*. Available online at https://comparecamp.com/shoplifting-statistics/. Accessed 11/21/2021.

software that identifies shoplifting in progress. The "VAAKeye" software analyzes store's surveillance cameras videos in real time. VAAKeye has been trained to identify 100 suspicious behaviors, including facial expressions, clothing choices, "restless behavior," hand movements, gait, and the sneaking of items into pockets and bags. When a suspicious behavior is detected, employees are alerted via a phone app. Tests show that VAAKeye reduces shoplifting by 77 percent.[68]

   a. Identify the type of technology used to make VAAKeye work. Be as specific as possible and explain how it works.
   b. What could cause the software to falsely identify someone as a shoplifter (who is not)?
   c. What additional technologies could be used to reduce shoplifting, especially at self-checkout stations?

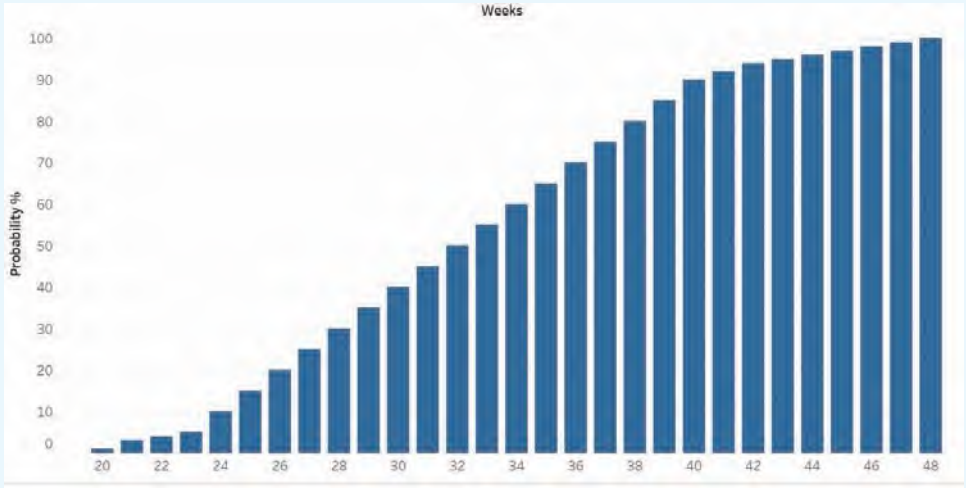
4. (LO 11.5) A CEO of a large furniture retailer wants to build a new intercompany accounts payable system. The CEO wants the system to record payment transactions between the company and its vendors/suppliers, and to provide a balance and the due dates of payments to be sent to vendors/suppliers for goods received. The company has worked with the vendors/suppliers for many years and has a good relationship with them. Due to security concerns, only certain members of the company will have access to this information. The CEO plans to use a blockchain to implement the system. Evaluate the CEO's decision to implement the system using a blockchain. What is your recommendation about developing the system using blockchain?
5. (LO 11.2, 11.3) Monte Carlo simulation estimates the probabilities of uncertain events. To execute a Monte Carlo simulation, we start by estimating the best-case, expected, and worst-case outcomes. Assume we are using Monte Carlo simulation to estimate the time to implement a real estate blockchain system with blockchain analytics and a user-friendly interface for 50 investors in a large condominium complex. The blockchain will keep track of all transactions, events, and documents related to the complex, including sales, purchases, titles, and credit checks. The system requires programming five different sections of code that work together. Given that blockchain is a newer technology, it is uncertain how many weeks will be needed to create each section of code. The following three estimates (A, B, and C) in weeks have been made to set parameters to run the simulation.

   a. What question are we answering with this Monte Carlo simulation?
   b. Identify which column (A, B, or C) is the best case, expected, and worst case. What is the total coding time estimated by each column?

Section	(A)	(B)	(C)
1	8	10	5
2	6	9	4
3	4	5	3
4	6	7	5
5	9	12	6

6. (LO 11.2, 11.3) Monte Carlo simulation estimates the probabilities of uncertain events. To execute a Monte Carlo simulation, we start by estimating the best-case, expected, and worst-case outcomes. Assume we are using Monte Carlo simulation to estimate the time to implement a real estate blockchain system with blockchain analytics and a user-friendly interface for 50 investors in a large condominium complex. The Blockchain will keep track of all transactions, events, and documents related to the complex, including sales, purchases, titles, and credit checks. The system requires programming five different sections of code that work together. Given that blockchain is a newer technology, it is uncertain how many weeks will be needed to create each section. The following Monte Carlo probability graph has been generated based on 1 million iterations.

[68]Source: Jaclyn Jeffrey-Wilensky, March 18, 2019, "Here's How AI Could Help Catch Shoplifters in the Act." *MACH.* Available online at https://www.nbcnews.com/mach/science/here-s-how-ai-could-help-catch-shoplifters-act-ncna984566. Accessed 11/21/2021.



a. What is the (estimated) minimum time required to create the blockchain system? What is the associated probability of completing the project in that number of weeks?

b. What is the (estimated) maximum time required to create the blockchain system? What is the associated probability of completing the project in that number of weeks?

c. If you wanted to be 80 percent certain that the project will be completed before or on a specified week when scheduling the project, how many days would you schedule the project for?

7. (LO 11.2, 11.3) A company created a cohort analysis, with cohorts defined as customers with an initial purchase in the same month. The company wants to know what the repurchase rate is in the months following the initial purchase. To answer this question, the company prepared the following chart for 2023.

Count of CustomerID	Column Labels												
Row Labels	0	1	2	3	4	5	6	7	8	9	10	11	Grand Total
⊞Jan	289	45	51	42	50	48	41	43	45	43	41	39	777
⊞Feb	194	31	31	31	19	28	30	26	36	24	16		466
⊞Mar	207	35	40	30	37	25	32	42	31	16			495
⊞Apr	170	20	22	25	21	30	18	15	13				334
⊞May	118	18	18	16	13	25	15	9					232
⊞Jun	109	14	15	18	15	14	12						197
⊞Jul	85	14	15	14	9	6							143
⊞Aug	74	11	14	16	11								126
⊞Sep	63	6	9	10									88
⊞Oct	43	7	3										53
⊞Nov	38	2											40
⊞Dec	48												48
Grand Total	1438	203	218	202	175	176	148	135	125	83	57	39	2999

a. How many total customers are tracked during the year?

b. How many purchases were made during the year?

c. How many customers made their initial purchase in December?

d. How many total purchases were made by the February cohort over the entire year?

e. What percentage of the customers repurchased items in month 1, the month after the first purchase? (Exclude the December cohort in the calculation.)

f. Should the company be concerned about its sales? Why or why not?

8. (LO 11.3, 11.4) Companies can use AI in many ways to reduce costs and/or increase revenues. For example, fast-food chains are incorporating AI into their operations to reduce customer wait times and to increase order accuracy with a goal of improving the overall customer experience, so that customers keep coming back. Below are some of the ways that companies are incorporating AI in operations. Answer the question posed about each scenario.
   a. **McDonald's** is testing new technology in its drive-thru lane. Cameras read cars' license plates and, based on prior orders, AI customizes the menu shown on a drive-thru menu screen. What type of advanced analytics is being used to read the license plates?
   b. McDonald's is also testing an additional technology. Chatbots will take orders from customers. What type of advanced analytics is being used to understand customers?
   c. **KFC** is allowing customers to take their picture with a machine that identifies the customer's age, gender, and even mood. Based on these characteristics, the AI system recommends menu items typically eaten by customers with similar traits. What type of advanced analytics is being used to recognize customers' traits?
   d. **Wendy's** First Kitchen locations in Japan are allowing customers to see their order history and pay with **PopPay**. PopPay links a customer's face with a credit card. What type of advanced analytics is being used to identify a customer's face?
   e. To increase profits, **Amazon** changes its prices on products every 10 minutes on average based on conditions, even changing prices between an item being placed in a cart and being paid for.[69] What AI application discussed in this chapter is Amazon using to identify the proper price for a product?
   f. According to a lawsuit filed in 2019, **Lowe's** and **Home Depot** have both added software to their surveillance cameras. The software tracks customers' movement through the store using a unique scan of face geometry. The footage is then shared with other stores. The stated goal of the faceprint collection is to reduce shoplifting and identify suspicious shopping behaviors. What type of advanced analytics is being used to recognize customers' faces?[70]



DATA VISUALIZATION

Data Visualization Exercise: Cycle Plot of Cohort Analysis

1. A company created a cohort analysis for its customers during the year. The company defined cohorts based on the month of initial purchase. Thus, there are 12 cohorts. The exhibit below shows the repurchase rates for each cohort during the year by month.

Count of CustomerID	Column Labels											
Row Labels	0	1	2	3	4	5	6	7	8	9	10	11 Grand Total
Jan	100%	16%	18%	15%	17%	17%	14%	15%	16%	15%	14%	13%
Feb	100%	16%	16%	16%	10%	14%	15%	13%	19%	12%	8%	
Mar	100%	17%	19%	14%	18%	12%	15%	20%	15%	8%		
Apr	100%	12%	13%	15%	12%	18%	11%	9%	8%			
May	100%	15%	15%	14%	11%	21%	13%	8%				
Jun	100%	13%	14%	17%	14%	13%	11%					
Jul	100%	16%	18%	16%	11%	7%						
Aug	100%	15%	19%	22%	15%							
Sep	100%	10%	14%	16%								
Oct	100%	16%	7%									
Nov	100%	5%										
Dec	100%											
Grand Total	100%	14%	15%	14%	12%	12%	10%	9%	9%	6%	4%	3%

[69]Source: Remi. "Companies That Use Dynamic Pricing." Available at https://www.remi.ai/post/companies-that-use-dynamic-pricing. Accessed on 12/19/2021.
[70]Source: Jonathan Bilyk, September 6. 2019, "Biometrics Class Actions Target Lowe's, Home Depot for Anti-Shoplifting Surveillance Systems." *Cook County Record.* Available online at https://cookcountyrecord.com/stories/513512142-biometrics-class-actions-target-lowe-s-home-depot-for-anti-shoplifting-surveillance-systems. 8/14/2022.

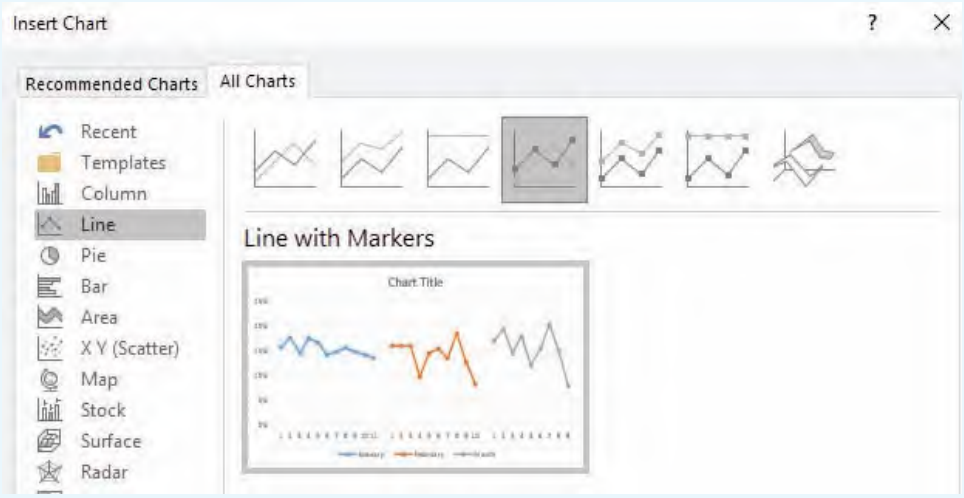
Suppose the company wants to easily see the trends in the data. Specifically, the company wants to answer the question "How do repurchase rate trends differ between the January, February, and March cohorts?" Sometimes it is easier to see trends in the data via a data visualization instead of a chart. For example, a *cycle plot* shows trends in data. Let's create a cycle plot for the first three cohorts (January, February, and March). The side-by-side nature of a cycle plot should help us see how the three cohort patterns/trends differ. To create the cycle plot, follow these steps:

1. Type the following data into Excel using the exact same format. To make the cycle plot work in Excel, the February cohort data need to start in a new column after the January data, and the March cohort data need to start in a new column after the February cohort data. In other words, the columns must be offset, as shown in the following exhibit, to make the cycle plot work in Excel.

	January	February	March
1	16%		
2	18%		
3	15%		
4	17%		
5	17%		
6	14%		
7	15%		
8	16%		
9	15%		
10	14%		
11	13%		
1		16%	
2		16%	
3		16%	
4		10%	
5		14%	
6		15%	
7		13%	
8		19%	
9		12%	
10		8%	
1			17%
2			19%
3			14%
4			18%
5			12%
6			15%
7			20%
8			15%
9			8%

2. Highlight the data.
3. From the menu, select **Insert > Charts > Recommended Charts.**

4. Select **All Charts > Line.**
5. Select the Line with Markers, which is the fourth chart from the left, as shown below:



6. Label the chart "Cohort Repurchase Rates".
7. Label the X-axis "Repurchase Month". *Hint:* **Chart Tools > Design > Add Chart Element > Axis Titles > Primary Horizontal.**
8. Label the Y-axis "Percent". *Hint:* **Chart Tools > Design > Add Chart Element > Axis Titles > Primary Vertical.**

CHAPTER 11 LABS

LAB 11.1 **Excel:** Advanced Analytics/Marketing: Analyzing Customer Repurchase Rates Using Cohorts

LAB 11.2 **Tableau:** Advanced Analytics/Marketing: Analyzing Text Using a Word Cloud

LAB 11.3 **Excel:** Advanced Analytics/Human Resources: Visualizing Survival Analysis

LAB 11.4 **Excel:** Advanced Analytics/Supply Chain: Estimating Demand Using Monte Carlo Simulation

LAB 11.5 **Advanced Analytics:** Understanding Blockchain

Lab 11.1 Excel

Advanced Analytics/Marketing: Analyzing Customer Repurchase Rates Using Cohorts[71]

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Cohort, Simulated Data, Cohort Analysis, Cohort Table, Retention Rates, Marketing

Decision-Making Context

This chapter discusses how business analysts can use analytics to dynamically segment the market in increasingly granular segments (of similar customers) to build fully customized products. One effective technique is cohort analysis. *Cohorts* are groups of people who share a common characteristic over time. Cohorts are used to tailor marketing efforts to customers with similar traits.

The fictitious Eagle Golf Company manufactures different types of customized golf products based on customer specifications. It has developed a new type of golf ball called MYOB (Make Your Own Ball). The balls are transparent and can have anything inside that the customer wants, such as flashing lights or team mascots. In this lab, we look at the future purchasing habits of a cohort, where the cohort is defined as customers who made their first purchase from Eagle Golf Company in the same month. Specifically, we will create a cohort analysis that shows what percentage of the cohort buys again from Eagle Golf during the rest of the year.

To help generate data for this lab, we used a website (http://www.databasetestdata.com/?source=post_page) that can help you create different types of data sets, which you can use to practice your analysis. Feel free to play around with the data you can create at this website.

[71]This lab is adapted from Yemi Johnson. "How to do your Cohorts analysis in Spreadsheet & Excel (A detailed guide)" Dec 27, 2016, Accessed: 8/14/2022. https://medium.com/@yemi_uc/how-to-do-your-cohorts-analysis-in-speadsheet-excel-a-detailed-guide-feddecb84c7a. Accessed on 8/14/2022.

Required

1. Create a cohort analysis for the rest of the year showing the repurchase rates at Eagle Golf Company.
2. Create a graph for the January cohort.

Specify the Question

How many members of a cohort purchase from Eagle Golf Company again during the year?

Obtain the Data

This lab uses the data in Lab 11.1 Data.xlsx. The data dictionary for this file follows.

Data Dictionary

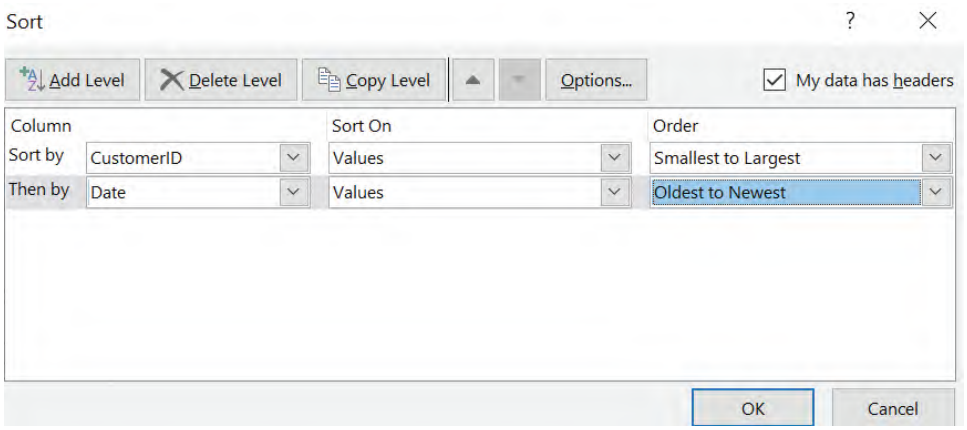
CustomerID:	Unique company-assigned customer number
Date:	Purchase date in MM/DD/YYYY format

Analyze the Data

Analysis Task 1: Set Up the Spreadsheet.

1. To sort the CustomerID in ascending order and the Date in ascending order as shown in Lab Exhibit 11.1.1, select **Data > Sort & Filter > Sort** from the menu. Then click OK.

Lab Exhibit 11.1.1



2. In cell C1, enter "Total Number of Sales Transactions".
3. In cell C2, enter the formula "=IF(A2=A1, C1+1,1)". Just like the COUNTIF function, which counts cells that match a given criterion, this formula counts the number of times that a customer buys from the company. However, the IF function is much more efficient (and therefore time saving) for large data sets than the COUNTIF function.
4. Copy the formula in cell C2 down the column. Lab Exhibit 11.1.2 shows what the data should look like. Note that customer 1 made three purchases while customer 2 made four purchases during the year.

Lab Exhibit 11.1.2

1	CustomerID	Date	Total Number of Sales Transactions
2	1	4/14/2022	1
3	1	5/26/2022	2
4	1	9/12/2022	3
5	2	5/1/2022	1
6	2	8/6/2022	2
7	2	8/8/2022	3
8	2	12/25/2022	4

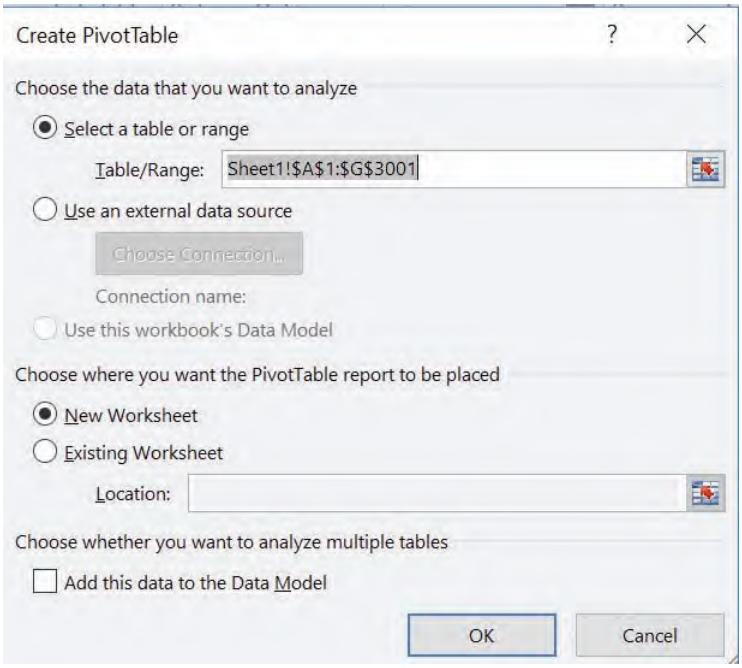
5. In cell D1, enter "Join Date". Join Date is the first time that a customer purchases from Eagle Golf Company.
6. In cell D2, enter the formula "=IF(C2=1,B2,D1)". (*Hint:* You may have to format the cell as a Date to make it show the proper information. This IF formula identifies the date of the first purchase and works because the data have been sorted first.)
7. Copy the formula down the column.
8. Eagle Golf Company creates cohorts based on the month of the first purchase. Therefore, in cell E1, enter "Cohort".
9. In cell E2, enter the formula "= MONTH(D2)". This function returns the numerical month in which the purchase took place (for example, January = 1, February = 2).
10. Copy the formula down the column.
11. In cell F1, enter "Month of Purchase".
12. In cell F2, enter the formula "=IF(AND(C2>1,MONTH(D2)=MONTH(B2)),MONTH(D2)+0.5,MONTH(B2))". The AND function checks whether all of the arguments listed are true. The AND function first checks if the total number of transactions is greater than 1, indicating the record is not the first purchase. The AND function then checks if the month of the purchase is the same as the month of the Join Date or first purchase. If all of the arguments are true, the AND function returns TRUE; if all of the arguments are not true, the AND function returns FALSE. If the AND function is true, then the IF function equals MONTH(D2)+0.5, or the number of the month of the Join Date + 0.5, indicating that the purchase was in the same month as the Join Date (or first purchase). IF the AND function is false, then the IF functions equals the Join Date month.
13. Copy the formula down the column.
14. In cell G1, enter "Number of Elapsed Months".
15. In cell G2, enter the formula "= F2-E2". This formula calculates the number of months between the first purchase and subsequent purchases. A 0.5 value indicates that the subsequent purchase was made during the same month of the initial purchase.
16. Copy the formula down the column.

Analysis Task 2: Create the Cohort Table.

We will use a pivot table to create a cohort table.

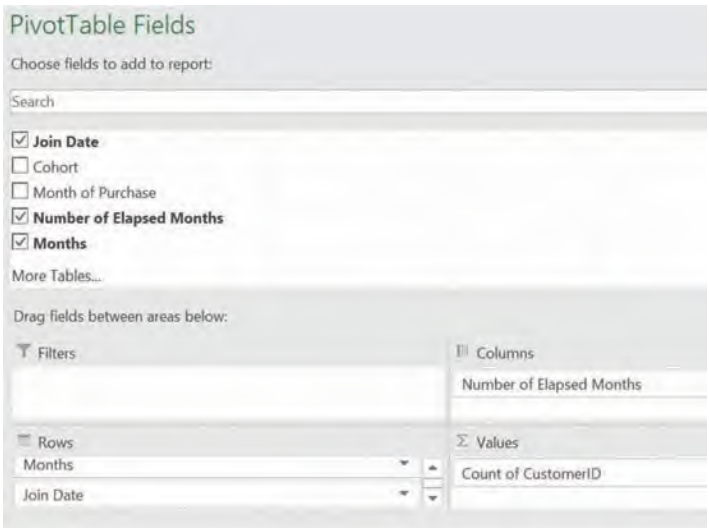
1. Place the cursor inside the data. From the menu, select **Insert > Tables > PivotTable**. The Create PivotTable window should appear and look like Lab Exhibit 11.1.3. Click OK.

Lab Exhibit 11.1.3



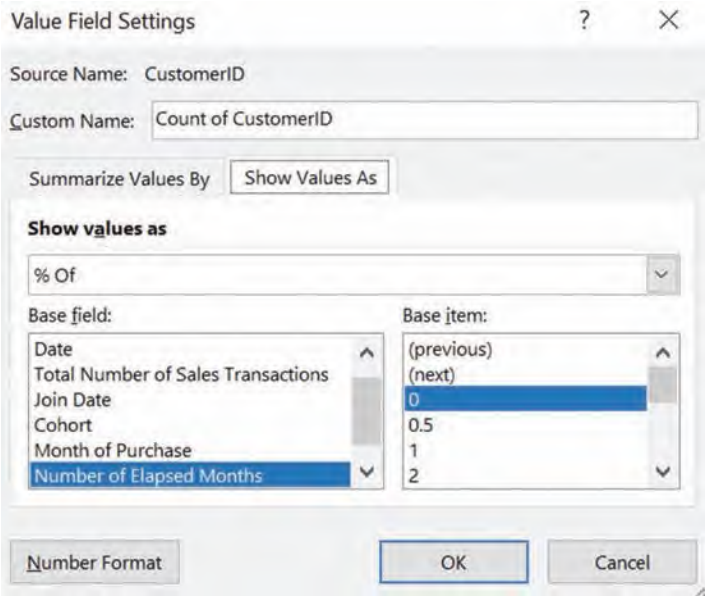
2. Click on the new worksheet. Drag Join Date to Rows. (*Note:* Months automatically shows up in Rows.) Drag CustomerID to ∑ Values. If the CustomerIDs are summed instead of counted, click on the down arrow by CustomerID. Select Value Field Settings. Select Count. Click OK. Drag Number of Elapsed Months to Columns. Lab Exhibit 11.1.4 shows how to set up the pivot table.

Lab Exhibit 11.1.4



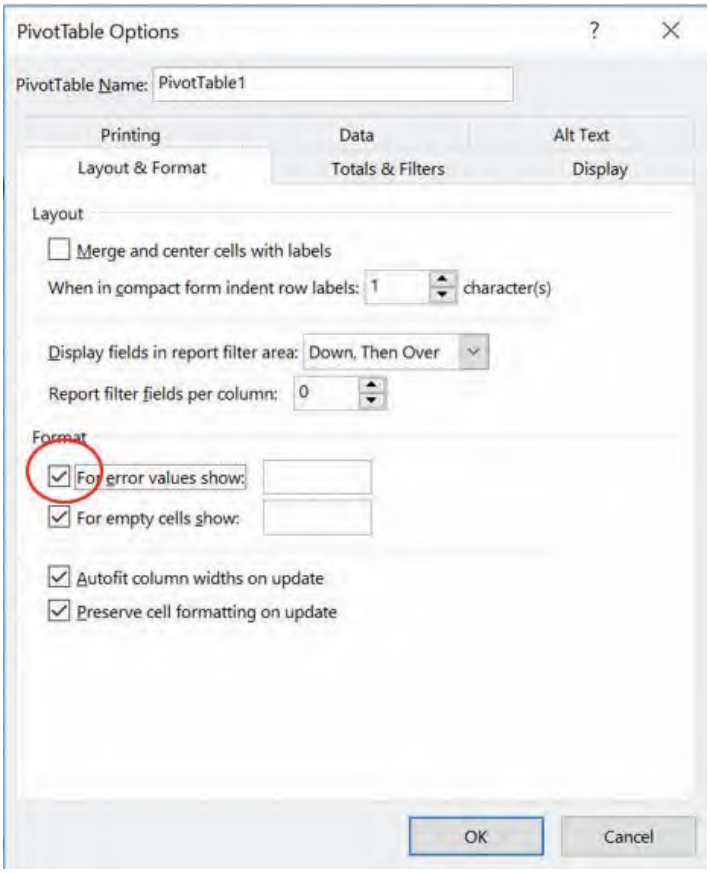
3. Label the worksheet "Count".
4. Copy the worksheet and label it "Repurchase Rates".
5. Now convert the totals to percentages. Click the arrow by Count of CustomerID. Select Value Field Settings. Click on the Show Values As tab. Select "% Of" with a Base field equal to "Number of Elapsed Months" and Base item equal to "0". Lab Exhibit 11.1.5 shows the Value Field Setting window. Click OK.

Lab Exhibit 11.1.5



6. To remove the #NULL! right-click inside the pivot table. Select PivotTable Options. Click the For error values show box, as shown in Lab Exhibit 11.1.6. The box should be empty. Click OK.

Lab Exhibit 11.1.6



Clicking on a cell in the pivot table makes Excel "drill down" to show the data points in the cell total. Drill-down is an important feature because it allows users to see data and information in more detail.



Report the Results

We have created Count and Repurchase Rates tables to see how many members of a cohort repurchase from Eagle Golf Company later in the year. For presentation purposes, let's create a data visualization of what we found for the first cohort, January.

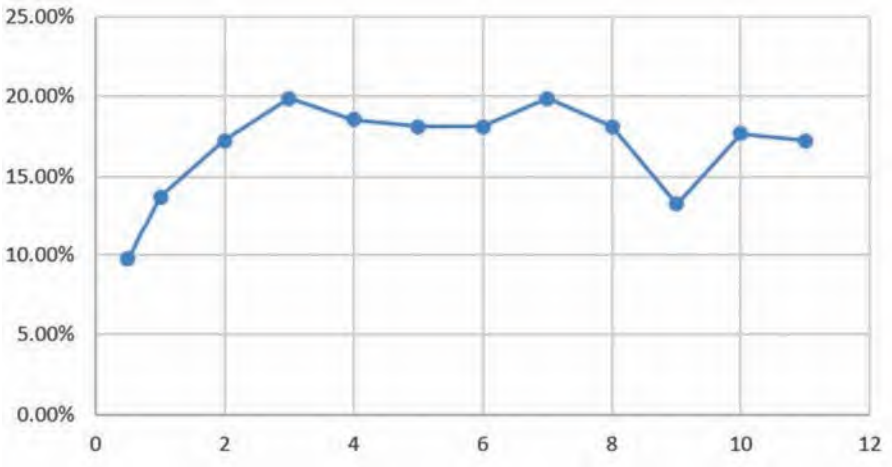
1. Copy the Repurchase Rates worksheet.
2. Label the worksheet "Graph".
3. Copy with a Transpose Paste the months 0.5-11 labels and the corresponding values for the January cohort. Note that the 100% is not included. Lab Exhibit 11.1.7 shows what the transposed data should look like.

Lab Exhibit 11.1.7

0.5	9.73%
1	13.72%
2	17.26%
3	19.91%
4	18.58%
5	18.14%
6	18.14%
7	19.91%
8	18.14%
9	13.27%
10	17.70%
11	17.26%

4. Highlight the data in Lab Exhibit 11.1.7.
5. From the menu, select **Insert > Charts > Recommended Charts.**
6. Select the first graph, called Scatter. Your graph should look like Lab Exhibit 11.1.8.

Lab Exhibit 11.1.8



7. Label the graph "January Cohort Repurchase Rate".
8. Label the X-axis "Month of Repurchase". (*Hint:* Use **Chart Tools > Design > Add Chart Element > Axis Titles > Primary Horizontal**.)

9. Label the Y-axis "Percent". (*Hint:* Use **Chart Tools > Design > Add Chart Element > Axis Titles > Primary Vertical**.)
10. Remove the decimals in the Y-axis. Right-click the Y-axis. Select Format Axis. Expand Number. Set Decimal places to 0.

Assessment

1. Take a screenshot of the completed Count table, copy it into a Word document named "Lab 11.1 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the Repurchase Rates table, copy it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the January repurchase graph, copy it into the same Word document, and label the screenshot Submission 3.
2. Answer the questions in Connect and upload your Lab 11.1 Excel submission.docx via Connect, if assigned.

Alternate Lab 11.1 Excel

Advanced Analytics/Marketing: Analyzing Customer Repurchase Rates Using Cohorts[72]

Apply the same steps in Lab 11.1 to the Alt Lab 11.1 Data.xlsx data set.

Required

1. Create a cohort analysis for the rest of the year showing the repurchase rates for Eagle Golf Company.
2. Create a graph for the January cohort.

Specify the Question

How many members of a cohort purchase from Eagle Golf Company again during the year?

Assessment

1. Take a screenshot of the completed Count table, copy it into a Word document named "Alt Lab 11.1 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the Repurchase Rates table, copy it into the same Word document, and label the screenshot Submission 2. Take a screenshot of the January repurchase graph, copy it into the same Word document, and label the screenshot Submission 3.
2. Answer the questions in Connect and upload your Alt Lab 11.1 Excel Submission.docx via Connect, if assigned.

[72]This lab is adapted from Yemi Johnson, "How to do your Cohorts analysis in Spreadsheet & Excel (A detailed guide)" Dec 27, 2016, Accessed: 8/14/2022. https://medium.com/@yemi_uc/how-to-do-your-cohorts-analysis-in-speadsheet-excel-a-detailed-guide-feddecb84c7a. Accessed on 8/14/2022.

Lab 11.2 Tableau

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Advanced Analytics/Marketing: Analyzing Text Using a Word Cloud

Keywords

Natural Language Processing, Word Cloud, Marketing

Decision-Making Context

Natural language processing (NLP) is a type of machine learning, and therefore AI, that helps computers understand, process, and generate human language. NLP can be used to analyze many types of text, including legal documents, social media posts, and financial statements. Google has an NLP Application Program Interface (API), or software that sends a user response to the computer system and then returns the system answer to the user, that can analyze text. To see how it works, visit https://cloud.google.com/natural-language/, scroll down to the Natural Language API demo, click Analyze, and check "I'm not a robot". Be sure to click on the Entities, Sentiment, Syntax, and Categories buttons to see what the analysis looks like. You can also try typing in text to see what the analysis looks like.

You can find another free website that can analyze text at https://voyant-tools.org/. To see how this website works, use the Lab 11.2 Amazon sample.xlsx data file. This file contains a sample of real reviews of **Amazon**'s Alexa. The data come from a post by Roshan Sharma in Kaggle.com, which is no longer available. Copy all the text in column D starting in row 2 to the Add text box at voyant-tools.org. Click Reveal. Voyant automatically analyzes the text. Among many other items, the analysis shows that there were 79,958 total words with an average of 13 words per sentence. If you click on Terms in the top-left corner, the results show that *love*, *alexa*, and *great* are the three most frequently used words in the reviews.

Although using NLP is beyond the scope of this book, we are going to look at one way to visualize the results of a textual analysis: a word cloud. A word cloud is a visual representation of text. Words that appear more frequently in the text are shown in larger font (or bolder colors). In this lab, we use Tableau to create a word cloud.

To see a video on how to generate a word cloud with Twitter data, visit https://www.youtube.com/watch?v=HDkI6G4slzQ. The video uses Orange, which is open-source (that is, free) machine learning and data visualization software, to analyze (or text-mine) Twitter. The end result is a word cloud, just like the one you will create in this lab. Many other data-mining software packages are also available, including **RapidMiner** and **KNIME**.

Magazines for All is a (fictitious) company that publishes many different magazines. *Outdoor Adventures* targets outdoor enthusiasts. To understand what readers would like to see in upcoming issues, Magazines for All asked the readers to post descriptions on social media of their favorite adventures.

Required

1. Create a word cloud and identify the word most frequently used.
2. Create a bar chart to identify the word most frequently used.

Specify the Question

What word is used the most frequently? This information will help editors understand what readers would like to see in upcoming issues of the magazine.

Obtain the Data

This lab uses the data in Lab 11.2 Data.xlsx. The file contains the results of a text analysis of readers' posts in social media. To generate a word cloud, we need a file with words

and usage rankings. In this lab, we rank the words using the percentage of responses including a particular description to describe a favorite adventure. The data dictionary for this file follows.

Data Dictionary

Description:	Word used to describe favorite adventure
Percentage:	Percentage of respondents using a particular description

Obtain Task 1: Load the Data into Tableau.

1. Open Tableau.
2. Connect to the Microsoft Excel file labeled Lab 11.2 Data.xlsx. Make sure that Description has been imported as text (Abc) and Percentage as number (#).

Analyze the Data

Analysis Task 1: Create the Word Cloud.

1. Click on Sheet 1.
2. We are going to focus on the Marks card to control how the data are displayed. Drag Description to Color. Drag Description to Label (it turns automatically into text). Drag Percentage to Size in the Marks card.
3. In the Marks card, change Automatic to Text using the drop-down arrow, as shown in Lab Exhibit 11.2.1.



Lab Exhibit 11.2.1

4. Name the sheet "Favorite Outdoor Adventures Word Cloud".

Report the Results: Create a Bar Chart

You have just used Tableau to create a word cloud describing the favorite adventures of *Outdoor Adventures* readers. The word cloud will help editors tailor future content to retain current subscribers and attract new readers. To help report the results to management, let's also create a bar chart for these data. Follow these steps:

DATA VISUALIZATION

1. Open a new worksheet.
2. Drag Description to the Columns shelf.

3. Drag Percentage to the Rows shelf.
4. Sort the bars in descending order.
5. Name the sheet "Favorite Outdoor Adventures Bar Chart".

Assessment

1. Take a screenshot of the word cloud, copy it into a Word document named "Lab 11.2 Tableau Submission.docx", and label the screenshot Submission 1. Take a screenshot of the bar chart, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Lab 11.2 Tableau Submission.docx via Connect, if assigned.

Alternate Lab 11.2 Tableau

Advanced Analytics/Marketing: Analyzing Text Using a Word Cloud

Magazines for All has another magazine called *Fitness Guru,* which inspires readers to lead a healthy life by presenting articles on different types of exercises, food, and motivation. Magazines for All asked its readers to use social media describe their favorite articles. Apply the same steps in Lab 11.2 to the Alt Lab 11.2 Data.xlsx data set. Instead of using percentage to rank the words, this lab uses total number of uses of a word. The data dictionary for the file follows.

Data Dictionary

Description:	Word used to describe favorite articles
Ranking:	Number of respondents using a particular description

Required

1. Create a word cloud and identify the word most frequently used.
2. Create a bar chart to identify the word most frequently used.

Specify the Question

What word is used the most frequently? This information will help editors understand what readers would like to see in upcoming issues of the magazine.

Assessment

1. Take a screenshot of the word cloud, copy it into a Word document named "Alt Lab 11.2 Tableau Submission.docx", and label the screenshot Submission 1. Take a screenshot of the bar chart, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Alt Lab 11.2 Tableau Submission.docx via Connect, if assigned.

Lab 11.3 Excel

Advanced Analytics/Human Resources: Visualizing Survival Analysis

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Keywords

Survival Analysis, Survival Curve, Human Resources

Decision-Making Context

Survival analysis estimates how the length of time before a particular event occurs. Although you may have heard about survival analysis in a medical context, in this lab we examine employee turnover at Eat-it-Fast, a (fictitious) local restaurant. Like all fast-food restaurants, Eat-it-Fast has trouble keeping employees for a long period of time. To help the restaurant plan, we will create a survival curve to analyze employee turnover data. A survival curve shows the fraction of the group surviving over time. A survival curve starts out at 100 percent at time zero and can decrease or remain the same from that time forward (but not increase). In our context, it shows the percentage of employees who remain with Eat-it-Fast over time. Companies can use this type of analysis as part of their HR analytics to predict turnover. This lab is based on https://www.real-statistics.com/survival-analysis/kaplan-meier-procedure/kaplan-meier-overview/.

Required

1. Create a survival curve for the employees at Eat-it-Fast.
2. Write a memo to summarize and explain your results.

Specify the Question

How many employees are still with Eat-it-Fast at a particular point in time?

Obtain the Data

This lab uses the data in Lab 11.3 Data.xlsx. Below is the data dictionary.

Data Dictionary

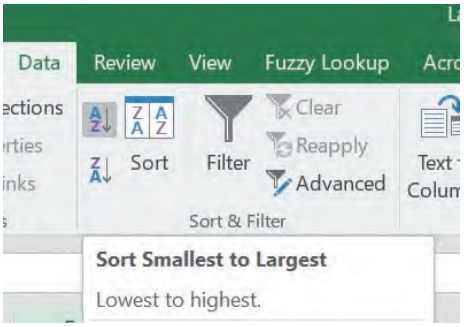
Months:	Number of months the employee was employed by Eat-it-Fast
Separated:	Equals 1 if employee no longer works for Eat-it-Fast; equals 0 if they are still working for Eat-it-Fast

Analyze the Data

Analysis Task 1: Calculate the Survival Function.

1. Copy column A to column D.
2. Change the title Months to "t" (to represent time).
3. Type "0" in cell D22 (this will be the start of the survival curve).
4. To remove duplicates in column D, place the cursor in the column and select **Data > Data Tools > Remove Duplicates** from the menu. Click OK. Click OK when Excel shows how many duplicates have been removed.
5. Then sort column D from smallest to largest. To do so, select **Data > Sort & FIlter > Sort Smallest to Largest** from the menu, as shown in Lab Exhibit 11.3.1.

Lab Exhibit 11.3.1



6. Enter the labels and formulas in Lab Exhibit 11.3.2 into the table. The variables and formulas used match the Kaplan-Meier procedure for survival analysis. Kaplan-Meier is a widely accepted method used to generate and analyze survival data. It is used to show how the data evolve over time.

Lab Exhibit 11.3.2

Cell	Enter	Meaning
E3	"#Quit"	Number of employees who quit at the point in time (t)
E5	=COUNTIFS($A$4:$A$21, D5,$B$4:$B$21,1)	Identifies observations in the MonthsEmployed column that match the value in the specified cell in the D column and then, for those observations, counts the number of 1s in the Quit column. This is the value of Quit at a particular point in time (t) identified in column D.
F3	"#Employed"	Number of employees still employed at Eat-it-Fast at the point in time (t) identified in column D.
F4	=COUNTIF($A$4:$A$21, ">"&D4-1)	Counts the number of observations in the Months column greater than the value in D4 minus 1 (or −1 for F4). This is the value of Employed at a particular point in time (t) identified in column D.
G3	"%Employed"	Percentage of survivors (in this case, employees still employed at Eat-it-Fast)
G5	=1-(E5/F5)	1 minus the ratio of Quit/Employed to determine the percentage of employed at a point in time (t) in column D
H3	S(t)	Survival function
H4	=1	All survival curves start at 1 (or 100%)
H5	=H4*G5	Survival function value

7. Copy the formulas in columns E, F, G, and H to row 16. After entering the information, the spreadsheet should look like Lab Exhibit 11.3.3.

Lab Exhibit 11.3.3

	A	B	C	D	E	F	G	H
1								
2								
3	Months	Separated		t	#Quit	#Employed	%Employed	S(t)
4	4	1		0		18		1
5	6	1		2	1	18	0.944444444	0.944444
6	7	1		4	1	17	0.941176471	0.888889
7	11	0		5	1	16	0.9375	0.833333
8	6	1		6	2	15	0.866666667	0.722222
9	6	0		7	2	12	0.833333333	0.601852
10	7	1		8	1	10	0.9	0.541667
11	11	1		9	0	9	1	0.541667
12	13	0		10	0	8	1	0.541667
13	14	0		11	3	7	0.571428571	0.309524
14	2	1		12	1	3	0.666666667	0.206349
15	11	1		13	0	2	1	0.206349
16	10	0		14	0	1	1	0.206349
17	9	0						
18	12	1						
19	5	1						
20	8	1						
21	11	1						

Analysis Task 2: Create the Survival Curve.

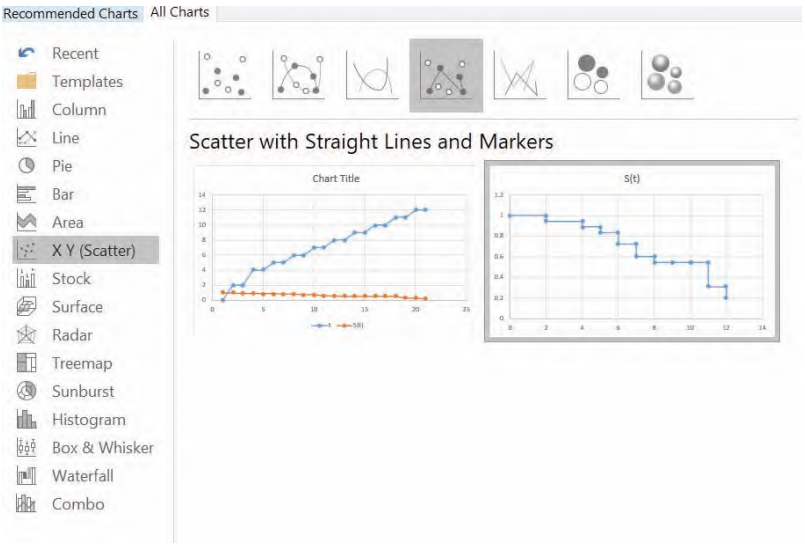
1. Copy the values in column D to column J.
2. Copy the values in column H to column K. To copy, first highlight the column, and then simultaneously hold down the ctrl key and the C key (ctrl-C). Release the keys. Then, move the cursor to the desired destination and right-click, selecting Paste Options Values (V).
3. Delete the rows where t = 13 and t = 14 because d = 0 for those two time measures (which means they don't add any new information). That is, delete cells J15:K16.
4. The survival curve is a stepwise curve due to the discrete nature of humans (we can't divide ourselves!). Therefore, we need to create the steps. Each step requires two values: a beginning value and an ending value. First, we need to duplicate the non-zero time values in column J below the original values. So, copy all non-zero values in cells J5:J14 to cells J15:J24. This creates the beginning point of the step (that is, the beginning or top of the riser). Next, we need to set the ending point of the step (or bottom of the riser). To do so, copy cells K4:K13 to cells K15:K24. Note that we are copying all values except the last one.
5. The last step before creating the curve is to copy the values in columns J and K to columns M and N. Then sort the columns first by t (smallest to largest) and then by S(t) (largest to smallest). The spreadsheet should look like Lab Exhibit 11.3.4.

Lab Exhibit 11.3.4

J	K	L	M	N
t	S(t)		t	S(t)
0	1		0	1
2	0.944444		2	1
4	0.888889		2	0.944444
5	0.833333		4	0.944444
6	0.722222		4	0.888889
7	0.601852		5	0.888889
8	0.541667		5	0.833333
9	0.541667		6	0.833333
10	0.541667		6	0.722222
11	0.309524		7	0.722222
12	0.206349		7	0.601852
2	1		8	0.601852
4	0.944444		8	0.541667
5	0.888889		9	0.541667
6	0.833333		9	0.541667
7	0.722222		10	0.541667
8	0.601852		10	0.541667
9	0.541667		11	0.541667
10	0.541667		11	0.309524
11	0.541667		12	0.309524
12	0.309524		12	0.206349

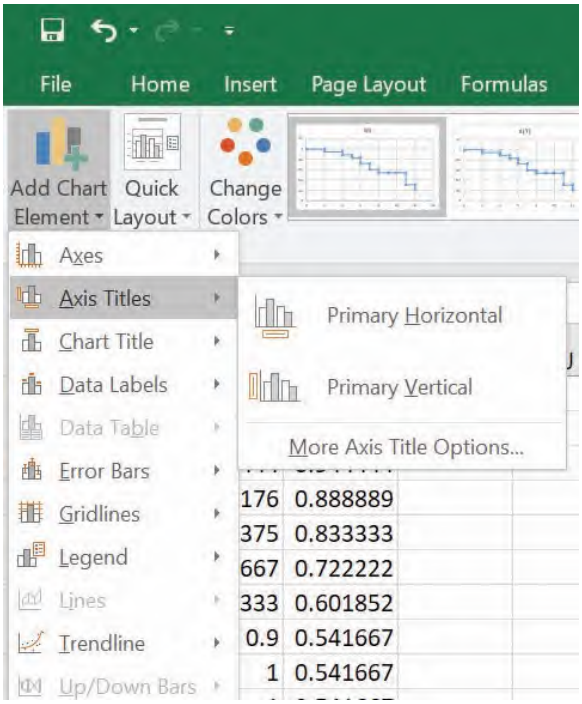
6. We can now use columns M and N, which represent a point in time (t) and the corresponding survival rate function value (S(t)), to create a survival curve. Highlight the data and headers in columns M and N. From the menu, select **Insert > Charts > Scatter with Straight Lines and Markers** as shown in Lab Exhibit 11.3.5. (*Hint:* Click on Recommended Charts, click on the All Charts tab, click on X Y (Scatter), click on the fourth chart from the left entitled Scatter with Straight Lines and Markers, click on the chart on the right, and click OK. The chart may need to be moved.)

Lab Exhibit 11.3.5



7. To make the chart more readable, click on S(t) and change to Employee Survival Curve. Add horizontal (Months) and vertical (Proportion) axis labels by selecting the **Add Chart Element > Axis Titles** option in the Chart Tools Design tab, as shown in Lab Exhibit 11.3.6. (*Hint:* Click on Primary Horizontal and type "Months" and then click on Primary Vertical and type "Proportion".)

Lab Exhibit 11.3.6



Analysis Task 3: Interpret the Survival Curve.

Here are a few important points about the survival curve:

- The horizontal axis (X-axis) gives the time in months after the start of the observation.
- The vertical axis (Y-axis) gives the proportion of employees still working for the company.
- The graph has a "staircase" appearance because the proportion does not change between employees leaving the restaurant.

Report the Results: Present Your Results

We have just created a Survival Curve to help management forecast employee needs. Write a memo explaining your results and what they mean to Eat-it-Fast.

Assessment

1. Take a screenshot of the survival curve, copy it into a Word document named "Lab 11.3 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the memo, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Lab 11.3 Excel Submission.docx file via Connect, if assigned.

Alternate Lab 11.3 Excel

Advanced Analytics/Human Resources: Visualizing Survival Analysis

Apply the same steps in Lab 11.3 to the Alt Lab 11.3 Data.xlsx data set.

Required

1. Create a survival curve for the employees at Eat-it-Fast.
2. Write a memo to summarize and explain your results.

Specify the Question

How many employees are still with Eat-it-Fast at a particular point in time?

Assessment

1. Take a screenshot of the survival curve, copy it into a Word document named "Alt Lab 11.3 Excel Submission.docx", and label the screenshot Submission 1. Take a screenshot of the memo, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Alt Lab 11.3 Excel Submission.docx file via Connect, if assigned.

Lab 11.4 Excel

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Advanced Analytics/Supply Chain: Estimating Demand Using Monte Carlo Simulation

Keywords

Monte Carlo Simulation, Supply Chain

Decision-Making Context

Monte Carlo simulation uses random sampling in conjunction with statistics to estimate mathematical functions to solve complex operations/systems. To use Monte Carlo simulation, we first select a distribution, repeatedly sample from the distribution, and then add up the results. Simply put, we use Monte Carlo simulation to estimate the probabilities of uncertain events. It is like flipping a coin a million times to see what happens. Companies use Monte Carlo simulations to make many decisions—for example, to hedge foreign exchange risk, postpone projects, decide on optimal investment strategies, estimate interest rate risk, and determine which products go to market.[73] Although advanced Monte Carlo simulations are beyond the scope of this textbook, this lab will introduce a simple Monte Carlo simulation.

The fictitious Eagle Golf Company manufactures different types of customized golf products based on customer specifications. It has developed a new type of golf ball called MYOB (Make Your Own Ball). The balls are transparent and can have anything inside that the customer wants, such as flashing lights or team mascots. Eagle Golf Company wants to calculate the expected total contribution margin that its MYOB golf ball will create. *Contribution margin* equals sales minus the variable cost of producing the items (or, on a per-unit basis, sales price per unit minus variable cost per unit). It is called contribution margin because it contributes to covering fixed costs. In the real world, we would be simulating many variables. In this lab, we will just simulate demand as it affects the total contribution margin.

Required

1. Use the RAND() function to simulate demand for a case of MYOB golf balls. Use the information to estimate the contribution margin from MYOB sales.
2. Create a bar chart of estimated demand levels.

Specify the Question

1. What is the expected demand for Eagle Golf's MYOB golf balls?
2. What is the expected total contribution margin for Eagle Golf's MYOB golf balls?

Obtain the Data

An Excel template to help complete the lab is provided in Lab 11.4 Data.xlsx. Unlike most other labs, this lab does not provide a data dictionary.

Analyze the Data

Analysis Task 1: Set Up the Spreadsheet.

1. Cells F4:G7 show the estimated demand and associated probability (based on market research). That information is used to create a table in cells I3:17 that can be used

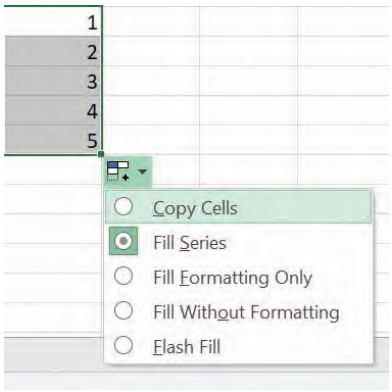
[73]Source: Introduction to Monte Carlo Simulation in Excel. Available online at https://support.office.com/en-us/article/Introduction-to-Monte-Carlo-simulation-in-Excel-64c0ba99-752a-4fa8-bbd3-4450d8db16f1. Accessed 11/22/2021.

in cells I3:17 that can be used in a VLOOKUP to convert the probability information into estimated demand. Cells G10 and G11 show the sales price and variable cost information on a per-unit basis (so the contribution margin per unit is $15). Lab Exhibit 11.4.1 shows the spreadsheet template. Enter information in each the cell highlighted in yellow, following the directions below.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Trial	Random Number	Estimated Demand	Estimated CM		Given Information:			Determine Cutoffs:			Outputs				
2																
3						Demand	Probability		Cutoffs	Demand		Trial mean				
4						5000	0.1		0	5000		Avg Estimated Demand				
5						15000	0.35		0.1	15000		Avg Estimated CM				
6						35000	0.3		0.45	35000						
7						60000	0.25		0.75	60000						
8												Demand	Fraction of Time			Contribution Margin
9												5000				
10						sales price	20					15000				
11						variable cost	5					35000				
12												60000				
13																
14												expected demand			expected CM	
15																

Lab Exhibit 11.4.1

2. In column A, enter the numbers 1 through 1000. (*Hint:* The easiest way to do this is to use the **Auto Fill Options > Fill Series** as shown in Lab Exhibit 11.4.2. These numbers identify our simulation trials, which means that 1,000 simulations will be run.)



Lab Exhibit 11.4.2

Analysis Task 2: Set Up and Run the Simulations.

1. In B2, enter the function "= RAND()". This function produces a number that is equally likely to assume any value between 0 and 1. Note that every time something changes in the spreadsheet, new numbers are generated. The F9 key also generates new numbers.
2. In C2, enter the function "=VLOOKUP(B2,$I$4:$J$7,2,TRUE)". This function looks up the demand based on the generated random number.
3. In D2, enter the formula "=C2*($G$10-$G$11)". This formula calculates the total estimated contribution margin based on the estimated demand.
4. Copy the contents of cells B2, C2, and D2 to row 1001.

Analysis Task 3: Analyze the Results.

To the right of the data, calculate the following items:

1. In cells M3, M4, and M5: Calculate the average of columns B, C, and D, respectively, using the Average function.

2. In cells M9–M12: Calculate the fraction/percentage of time that demand equaled each of the four given demand values in F4, F5, F6, and F7 (5000, 15000, 35000, 60000) using the COUNTIF function. (*Hint:* In cell M9, enter the formula "=COUNTIF($C$2:$C$1001,L9)/1000)", which counts the number of cells between C2 and C1001 that equal the value of cell L9, or 5000 in this problem.)
3. In cell M14: Calculate the expected demand, which is the sum of the demand value times the fraction of time for the associated demand value. So, enter the formula "=L9*M9+L10*M10+L11*M11+L12*M12".
4. In cells P9–P12: Calculate the total contribution margin for each demand value. Total contribution margin equals demand * (sales price – variable cost). (*Hint:* In cell P9, enter the formula "=L9*($G$10-$G$11)".)
5. In cell P14: Calculate the expected total contribution margin, which is the sum of the contribution margin times the fraction of time for the associated demand value. So, enter the formula "=M9*P9+M10*P10+M11*P11+M12*P12".

Report the Results: Bar Chart of Expected Demand

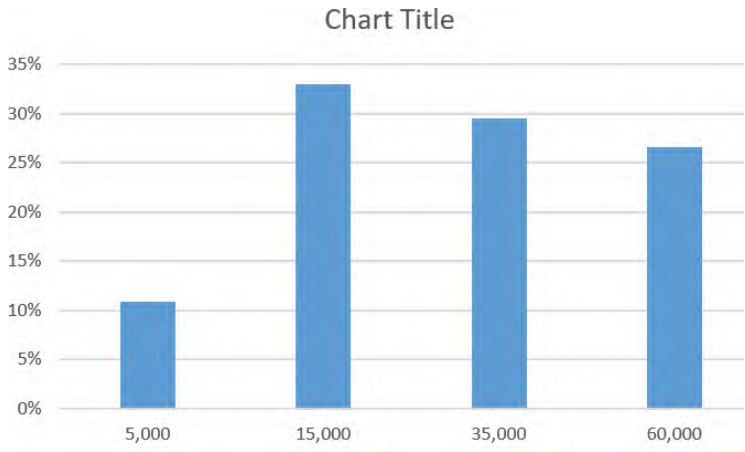


DATA VISUALIZATION

We have just used a Monte Carlo simulation to estimate demand and projected contribution margin for a new product, MYOB golf balls. The projected demand may also help managers plan production schedules and marketing campaigns. To help managers understand the results of the simulation, let's create a bar chart of the four demand levels and the corresponding probabilities by following the steps below.

1. Highlight cells L9:M12.
2. From the menu, select **Insert > Charts > Recommended Charts.** Select the chart called Clustered Column. Your choice should look like Lab Exhibit 11.4.3, with Demand on the X-axis and fraction of time on the Y-axis.

Lab Exhibit 11.4.3



Now let's make the chart easier to read.

3. Label the chart "Monte Carlo Simulation Demand Estimate".
4. Label the X-axis "Demand".
5. Label the Y-axis "Fraction of Time".

Assessment

1. Take a screenshot of the first 20 rows and all the columns of the spreadsheet, copy it to a Word document, and label it "Lab 11.4 Excel Submission.docx". Take a screenshot of the bar chart, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Lab 11.4 Excel Submission.docx file via Connect, if assigned.

Alternate Lab 11.4 Excel

Advanced Analytics/Supply Chain: Estimating Demand Using Monte Carlo Simulation

Apply the same steps in Lab 11.4 to the Alt Lab 11.4 Data.xlsx data set, which has the following changes:

Sales price = 30
Variable cost = 6.50

Demand	Probability
3,500	0.10
7,000	0.15
10,000	0.27
20,000	0.35
35,000	0.10
57,000	0.03

Required

1. Use the RAND() function to simulate demand for a case of MYOB golf balls. Use the information to estimate the contribution margin from MYOB sales.
2. Create a bar chart of estimated demand levels.

Specify the Question

1. What is the expected demand for Eagle Golf's MYOB golf balls?
2. What is expected total contribution margin for Eagle Golf's MYOB golf balls?

Assessment

1. Take a screenshot of the first 20 rows and all the columns of the spreadsheet, copy it to a Word document, and label it "Alt Lab 11.4 Excel Submission.docx". Take a screenshot of the bar chart, copy it into the same Word document, and label the screenshot Submission 2.
2. Answer the questions in Connect and upload your Alt Lab 11.4 Excel Submission.docx file via Connect, if assigned.

Lab 11.5

Lab Note: The tools presented in this lab periodically change. Updated instructions, if applicable, can be found in the student and instructor support materials.

Advanced Analytics: Understanding Blockchain

Keyword

Blockchain

Decision-Making Context

One trending technology is blockchain, a digital, distributed ledger that stores data between entities in real time and is very difficult to change. Blockchain consists of records or blocks (hence, the block part) in which records/blocks are stored in a continuous linked chain (hence, the chain part). Blocks contain data about something. For example, a block may record a transaction that took place between a supplier and a company that purchased inventory from the supplier. Each transaction could be a different block in the chain in chronological order. The collection of blocks is sometimes called a digital ledger. To get access to the blockchain, the blockchain members interact via a computer network connection to share information. All of the blockchain members, who are called peers, have access to the entire blockchain, and each peer stores its own copy, which is what makes the ledger "distributed."

The first block in the chain is called the *genesis block*. To add additional blocks, blockchain outsiders, called *miners,* solve complex math problems to verify that the block is valid in exchange for some form of compensation, such as cryptocurrency or information. This mining process is called *proof-of-work*. (Some blockchains use a different approach than mining, but with the same verification purpose.) Once the block is verified, it is added to the blockchain using a cryptographic hash. Cryptography uses complex mathematical codes to store and transmit data between entities with the goal of ensuring that the data can be read only by authorized participants.

While there are many techniques in cryptography, blockchain uses a *hash,* which is a function that takes data and turns it into a digital fingerprint of data with a standard (agreed-upon) length. Each block has a unique hash that is a combination of letters and numbers. So, if you change any of the data in a block, the hash changes. In addition to having a hash, each block contains the hash of the prior block, creating a chain. If information is changed in a block, then the hash changes, making it not match the original hash and breaking the chain. The use of hashes makes it difficult to change the data because any data changes break the blockchain hashes. This property of blockchains is often called *immutability*. The goal of this lab is to help you understand how blockchains work.

Required

1. Explore how blockchains work.
2. Write a memo describing how blockchains work.

Specify the Question

How do blockchains work?

Obtain the Data

There are no data files to use in this lab. Instead, go to the following website: https://andersbrownworth.com/blockchain/. This website is designed to help you understand how blockchains work. You will watch a video about blockchains and then perform some steps to better understand what is going on. Questions are posed in the directions to help you understand how blockchain works.

Analyze the Data

Analysis Task 1: Watch Blockchain 101—A Visual Demo.

1. Watch Blockchain 101—A Visual Demo on the website. The SHA256 is a cryptographic algorithm (or function) that generates a unique, fixed 256-bit-long hash for data with a length of 64 characters. The *nonce* is the (seeded) value needed to obtain a specific hash value. This is the value calculated by miners to earn compensation. A green block means that the block is validated or signed. A red block means that the block is not validated or signed and needs to be re-mined.

   Other terms introduced in the video include tokens and coinbase. A *token* is a digital resource that a user controls and can reassign to someone else. It can be any asset, including cryptocurrency, bonds, stocks, and shares of a company's stock. *Coinbase* is essentially a wallet for digital currencies. We will not focus on these two aspects of blockchain. Instead, the lab will focus on understanding the concepts of hash, block, blockchain, and distributed.

Analysis Task 2: Begin Your Exploration with Hash.

1. Click on the Hash tab (https://andersbrownworth.com/blockchain/hash), which allows you to see how a basic hash function works. The SHA256 Hash box should appear.
2. Type your college/university name in the Data box first in all lowercase letters and then a second time in all uppercase letters. How often does the hash change? Is the hash case-sensitive?

Analysis Task 3: Continue Your Exploration with Blockchain.

1. Click on the Block tab (https://andersbrownworth.com/blockchain/block), which shows you how a block might look. In this example, the block consists of four elements: Block (number), Nonce, Data, and Hash. The Block number indicates the order of the blocks. The Nonce is solved by the Mine button (acting as a blockchain miner) to obtain a particular Hash. Data is the information of interest. Hash is the digital fingerprint of all the information in the block. So, if any element in the block changes, so does the Hash. Remember, a green block means that the block is validated or signed. A red block means that the block is not validated or signed and needs to be re-mined.
2. As you complete this step as well as the following steps, keep track of the beginning and ending values of the hash. Type your college/university name in the Data block. What happens? What needs to be done? How does the hash change?

Analysis Task 4: Continue Your Exploration with Blockchain.

1. Click on the Blockchain tab (https://andersbrownworth.com/blockchain/blockchain), which shows how a chain of blocks is maintained. Prev has been added as an additional field. What is Prev?
2. Scroll right to see how many blocks are in the chain.
3. Type your college/university name in the last block. What happens? What needs to be done to fix any problems? Perform the step(s) to fix the block.
4. Now, type your college/university name in the third block. What happens? What needs to be done to fix any problems? Perform the step(s) to fix the block.

Analysis Task 5: Continue Your Exploration with the Distribute Tab.

1. Click on the Distributed tab (https://andersbrownworth.com/blockchain/distributed). The Distributed tab shows how a distributed blockchain works. A blockchain is distributed when more than one entity (or peer) has a copy of the blockchain.
2. Scroll down to see how many peers have a copy of the blockchain.
3. In Block 2 for Peer B, type your college/university name. What happens? If you fix the problem like in the Blockchain tab, is the problem solved? How could other members of the blockchain be notified of a problem?

Report the Results: Write a Memo on How Blockchain Works

You used a website to understand how blockchain works. Blockchains are growing in usage, making it critical for you to have a basic working knowledge of the technology. Write a business memo in proper format describing how blockchain works.

Assessment

1. Save your memo in a Word document labeled "Lab 11.5 Blockchain Submission. docx".
2. Answer the questions in Connect and upload your Lab 11.5 Blockchain Submission. docx file via Connect, if assigned.

Alternate Lab 11.5

Advanced Analytics: Understanding Blockchain

Watch the Blockchain Demo–Part 2–Public/Private Keys & Signing.

Required

Explore how blockchains work.

Specify the Question

How do blockchains work in general and, specifically, what is the role of public and private keys?

Obtain the Data

There are no data files to use in this lab. Instead, go to the following website: https://andersbrownworth.com/blockchain/.

Assessment

Answer the questions in Connect after reviewing the Chapter 11 and Lab 11.5 materials. The Connect questions cover all of the blockchain materials.

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Chapter 12

Using the SOAR Analytics Model to Put It All Together

Three Capstone Projects

A Look Back

- Chapter 11 described advances in technology being applied to business analytics to create a fifth type of analytics, adaptive/autonomous analytics.
- Adaptive/autonomous analytics takes analytics to a new level as reiterative technology executes decisions, scans the environment, adapts to changes, and then executes new decisions.

A Look at This Chapter

- This chapter offers three projects that allow you to put the SOAR analytics model into practice, from specifying the appropriate question all the way through reporting the results.
- The first project considers the determinants of Airbnb nightly rental rates in New York City.
- The second project considers the factors that may predict loan repayment at Lending Club.
- The third project provides an outline for completing a business analytics project of your choice, addressing a question in your specific area of interest and expertise.

Using Data Science to Assist in the Response to COVID-19

Throughout the COVID-19 pandemic, the **Center for Systems Science and Engineering** at Johns Hopkins University used multiple formats to display confirmed global cases and global deaths by country, region, and state. Often, the data were presented in a dashboard format, as in the accompanying illustration, which provides at-a-glance, color-coded information, using red for total cases, white for total deaths, and green for total vaccination doses administered. The World Health Organization, Centers for Disease Control, and local and country governments around the world will learn from these data for decades to come.

Why were some countries able to prevent the spread of the virus and others unable to do so? Which factors predicted this virus spread in each country, and how can we learn from this information to prevent the spread of future pandemics?



COVID-19 Dashboard

Source: https://www.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6. Accessed 2/02/22.

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

- **LO 12.1** Propose the appropriate price to charge for Airbnb rentals using the SOAR analytics model.
- **LO 12.2** Determine the factors associated with loan repayment by using the SOAR analytics model.
- **LO 12.3** Propose your own business-related question and answer the question using the SOAR analytics model.

Throughout this text, we've discussed the business analytics process using the SOAR analytics model (summarized in Exhibit 12.1). Recall that the four components of the SOAR model are:

1. **S**pecify the question (Chapter 1): Ask a question that is answerable with data.
2. **O**btain the data (Chapter 2): Gather applicable data to address the question.
3. **A**nalyze the data (Chapters 3–5 and 7–11): Analyze the data through various data analytics techniques.
4. **R**eport the results (Chapter 6): Communicate the findings to decision-makers and other interested parties.

It's now your turn to use the SOAR model, from start to finish, for one or more of the projects proposed in this chapter. The first project asks you to predict the rental rate for each night in an Airbnb rental unit in New York City. Using a database made available over the past few years, you'll use your data analytics skills to predict price. In the second project, you'll assess the factors associated with paying off a loan. Using a database of loan repayment from

Exhibit 12.1
The SOAR Analytics Model



Lending Club, you'll use your business analytics skills to determine which factors are important in predicting loan repayment. In the final project, which is more open-ended, you specify the relevant question and then use your business analytics skills to obtain the data, analyze the data, and report the results to answer a question that you have posed.

LO 12.1

Propose the appropriate price to charge for Airbnb rentals using the SOAR analytics model.

12.1 PROJECT 1: DETERMINING THE PRICE OF AIRBNB NIGHTLY RENTALS IN NEW YORK CITY

Airbnb (Exhibit 12.2) is an online marketplace that connects people who would like to rent an available room (or home) from someone who owns a room or home. Airbnb is the broker, or intermediary, that matches those wishing to rent with those who have the accommodations. Airbnb offers listings in over 81,000 cities in 191 countries and has about 150 million users.[1]

Exhibit 12.2



Daniel Krason/Shutterstock

In this assignment, we do not give you the same level of detailed lab instructions found in the rest of this textbook. Instead, we provide general directions. From there, you decide what needs to be done and the best ways to perform the analysis.

[1]https://news.airbnb.com/about-us/. Accessed 2/01/22.

Specify the Question

There are many potential questions, but the general question we want to answer is "What predicts the price for a nightly Airbnb rental in New York City?"

Imagine that you're a homeowner in New York City and you are trying to decide how to set prices for a nightly rental. Which factors are important? Which are not? What should you include in your model to predict the nightly rental price? Here are some factors you might consider:

- Location: Is it close to Central Park or the Empire State Building? Is it in Manhattan or the Bronx?
- Size: How large (or small) is the rental until?
- Furnishings: Are they old and dilapidated, or shiny and new?
- Privacy: Do you need to share the space with the renter, or will the renter have complete privacy?
- Reviews: If past reviews are strong and the rental unit is popular, can you charge more for it?
- Day of the week: Can you charge more on weekends than on weekdays?

Obtain the Data

Now that we've specified the question, it is time to consider which data are available to help us perform the analysis.

The data for this analysis are available at kaggle.com at this website (Exhibit 12.3): https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data. Kaggle, which is a subsidiary of Google, is an online community of data scientists and machine learning practitioners. It allows users to publish data sets, build models, and work with other members of the community.

Browse the site to get some ideas of what you can do with the data. There is plenty of information about the sample, much of it relevant to the question we have specified.

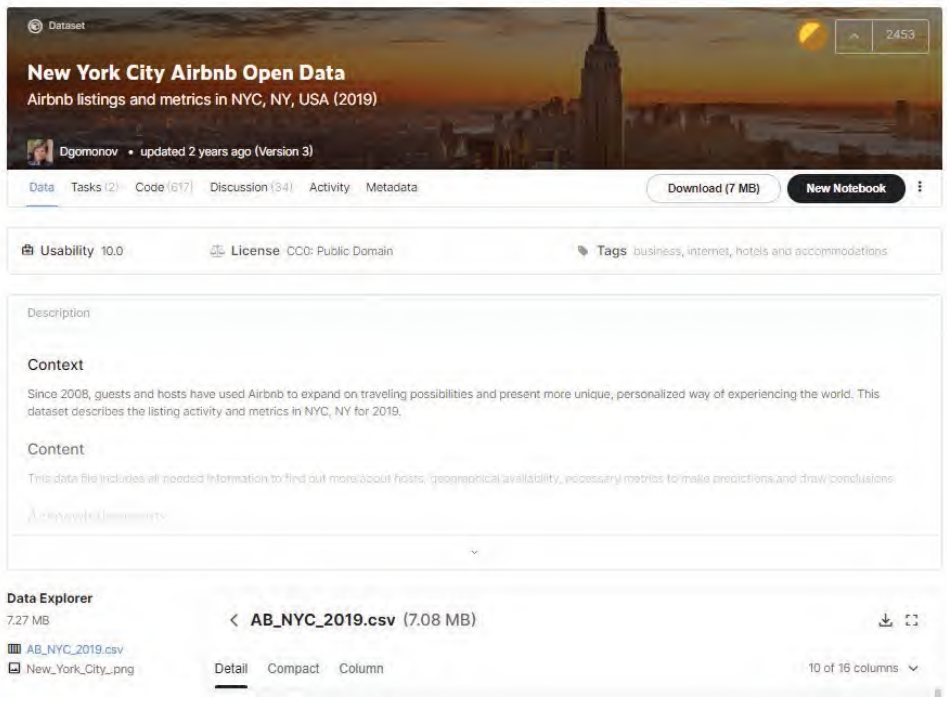


Exhibit 12.3 Kaggle Site with the New York City Airbnb Open Data

Source: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data. Accessed 8/2/2022.

Step 1: Download the Data and Review the Data Set

Download and open the file named "AB_NYC_2019.csv" in Excel and browse the data. Exhibit 12.4 is the data dictionary describing the data. (You may need to sign in using a Google account and agree to Kaggle's terms of use.)

Exhibit 12.4 Data Dictionary for Airbnb Listing Data in New York City

(Source of Data: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data. Accessed 8/2/2022.

Column	Variable Name	Variable Explanation
A	id	listing id
B	name	name of the listing
C	host_id	host id
D	host_name	name of the host
E	neighbourhood_group	broad location
F	neighbourhood	more specific neighborhood
G	latitude	latitude coordinates
H	longitude	longitude coordinates
I	room_type	listing space type (entire house, private room, or shared room)
J	price	price in dollars
K	minimum_nights	minimum rental nights required
L	number_of_reviews	how many reviews are made?
M	last_review	date of most recent review for this listing
N	reviews_per_month	reviews made within the last month
O	calculated_host_listings_count	count of host listings
P	availability_365	number of days available for rental per year

The Kaggle website contains additional descriptions, definitions, and discussions about the data and other types of analysis that could be performed on the data. To access this discussion, click on the "Discussion" tab.

Step 2: Identify the Dependent (Outcome) Variable

Because we are trying to determine the appropriate nightly rental rate, price is our dependent (outcome) variable.

Step 3: Identify Potential Independent (or Explanatory) Variables

The first step in determining the best data to use is to consider the various explanatory, or independent, variables. Which variables are most likely to explain the price charged for an Airbnb rental in New York City? Recall that the price of one night's rental is the dependent (outcome) variable that we are trying to explain (or predict). As an analyst, it is your job to identify the variables that are most likely to explain the price. Look over the data dictionary to determine which are likely to be the best candidate explanatory variables.

Deliverable 1:

Once you have identified the variables that may explain the price of an Airbnb rental, prepare a document (in Microsoft Word) and save it as "Project 1, Deliverable 1: Airbnb Price Analysis First Name and Last Name" (inserting your first and last name). Title the first section of the document "Dependent and Independent Variables". Next, within the document, identify and type the names of the dependent (outcome) variable and at least five proposed independent variables that you think explain price.

Step 4: Clean and Transform the Data in Preparation for Analysis

To prepare the data for analysis, complete this four-step process:

1. Because some of the listings have no date of last review, they are missing data. To remedy this problem and still be able to include the listing in the analysis, replace date of last_review with March 28, 2011. Color the font in **red** so you know you inserted those data.
2. There are also missing data for reviews_per_month. To remedy this problem, insert the value of zero (0) if reviews_per_month is missing. Color the font in **red** so you know you inserted those data.
3. Calculate a variable called EntDum (a dummy variable that can carry the value of 1 or 0, which equals 1 if "room_type" = "Entire House" and 0 if "room_type" = "shared room" or "private room"). To do so, sort the database by room_type and insert a variable and name it "EntDum".
4. The borough of Manhattan in New York is an important area for tourism, and it may well command a higher price for Airbnb rentals than accommodations in the other four boroughs (Brooklyn, Queens, Staten Island, and the Bronx). Create a new Manhattan dummy variable (a dummy variable that can carry the value of 1 or 0) that equals 1 if neighborhood_group = "Manhattan" and zero (0) otherwise.

Deliverable 2:

In the second section of your report (after Deliverable 1), prepare a descriptive analysis titled "Descriptives" that includes a screenshot of descriptive statistics for all the proposed dependent (outcome) and independent variables. Using the Descriptive Statistics tool in Excel's Data Analysis ToolPak (Exhibit 12.5), describe any unusual findings.

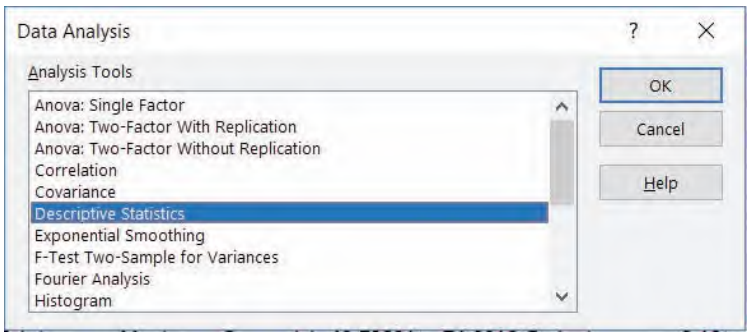


Exhibit 12.5
Descriptive Statistics in the Data Analysis ToolPak

Deliverable 3:

In the third section of your report (after Deliverable 2), predict how the independent variables might relate to the price charged for a one-night rental. For example, we would predict that the houses with more reviews are more popular and thus likely to command a higher price, so there would be a positive relationship between number of reviews and rental price. Make a prediction for each independent variable you think might be important and include them in your report as Deliverable 3.

Analyze the Data

Now that you have obtained the data, you can begin your analysis.

Step 1: Hypothesize Links Between Dependent/Outcome Variable (Price) and the Independent Variables You Chose

What do you hypothesize are the links between the dependent (outcome) variable of price and the independent variables that might predict it? Look through the list of variables in the data dictionary. Which ones predict the price of Airbnb rentals, and are they positively or negatively related to price?

Step 2: Select and Run Desired Analytics Techniques to Evaluate Relationships

What types of analysis will help show a relationship between price and the various independent variables? Consider the following:

- *Pivot tables:* Create a pivot table of price (dependent variable) and each independent variable you proposed (**Excel: Insert > Pivot Table**). After running the pivot analysis, do you still believe there is a relationship between the dependent variable and each independent variable you proposed?
- *Correlations:* Using the Data Analysis ToolPak (**Excel: Data > Analysis > Data Analysis > Correlation**), run a correlation between price (dependent variable) and a few independent (now transformed) variables one by one. After running the correlation analysis, do you still believe there is a relationship between the dependent variable and each independent variable you proposed?
- *Regression analysis:* Using the Data Analysis ToolPak (**Excel: Data > Analysis > Data Analysis > Regression**), run a regression of price (dependent variable) on your proposed independent variables. (For additional description of regression analysis, see the statistics review in Chapter 3.)



SOARing to Success

Data Tip

Recall that Excel requires you to group the independent variables together in side-by-side columns before you run a regression.

Deliverable 4:

In the fourth section of your report after Deliverable 3, prepare an analysis titled "Analysis" that includes the following:

1. Screenshots of at least two pivot table analyses and your interpretation (at least 100 words).
2. Screenshots of the output of at least two correlations between the dependent variable (price) and one independent variable. What is your interpretation of the findings?
3. Screenshot of the results (output) of the regression, including the dependent variable (price) and all of your proposed independent variables. What is your interpretation of the findings? (*Hint:* In general, if the *t*-statistic is greater than 2 or the *p*-value is less than 0.05, then the independent variable is likely a significant variable in predicting the price of a night's rental on Airbnb.)

Was the direction of the relationship for each variable (positive vs. negative vs. no relationship) different from what you anticipated in Deliverable 3?

Report the Results

Based on your analysis, consider how to present your findings to Airbnb management. Would you use graphs, tables, correlations, pivot tables, and/or regression results to share the story? How would you most effectively communicate the findings to Airbnb management?

Deliverable 5:

In the fifth section of your report after Deliverable 4, prepare an analysis titled "Final Report" that includes the following:

1. An executive summary that is a maximum of one page. It should concisely describe the variables that affect rental price per night and help management determine how to price Airbnb rentals. Be sure to provide the main support for your recommendation.
2. A report summarizing your findings and any graphs, reports, correlations, pivot tables, or regression results that help communicate your findings effectively.
3. A brief discussion of how these findings might help predict the price of a night's rental on Airbnb in the future even before an accommodation is listed.

12.2 PROJECT 2: DETERMINING THE FACTORS ASSOCIATED WITH LOAN REPAYMENT

LO 12.2

Determine the factors associated with loan repayment by using the SOAR analytics model.

Lending Club is a U.S.-based peer-to-peer lending company (www.lendingclub.com), headquartered in San Francisco, California. Lending Club facilitates both borrowing and lending by providing a platform for unsecured personal loans between $1,000 and $35,000. The loan period is either three years or five years. Potential investors search and browse the loan listings on the Lending Club website and select loans in which they would like to invest. The available information includes background about the borrower, amount of the loan, loan grade and related loan interest rate, and loan purpose. Investors finance the loans and make money from the interest paid by the borrower. Lending Club makes money by charging borrowers an origination fee and investors a service fee.

Imagine you are the accounting data analyst for Lending Club. You are given the task of developing a diagnostic model of loan repayment to facilitate the loan approval process. Using a Lending Club data set and the SOAR model, prepare and clean the data, choose appropriate explanatory variables, run appropriate analysis, and report the results to management. By more clearly understanding the link between initial loan approvals and ultimate loan repayment, Lending Club will be able to make more intelligent loan-approval decisions.

Specify the Question

What borrower or loan characteristics are associated with loan repayment?

Answering this question will help Lending Club determine who should or should not receive a loan in the future. It may also help predict the risk associated with a loan, which will influence the interest rate charged to the borrower (because, all other things equal, the higher the risk of repayment, the higher the interest rate the borrower pays).

Obtain the Data

The data for this analysis come from Lending Club and are available on the instructor website. Lending Club provides a summary data set of the loans extended to borrowers since 2007.

Step 1: Download the Data and Review the Data Set

The name of the available file is LoanStats3a.xlsx. It covers the repayment of loans made over a five-year period.

Open the file named "LoanStats3a.xlsx" in Excel, and look at the data. Each row represents a different loan. Each column represents an attribute of the loan or of the borrower. For example, Column D—funded_amnt—lists the amount of each loan that was funded.

Step 2: Identify the Dependent (Outcome) Variable

In column Q, titled "loan_status," there are two possible entries: "Charged Off" or "Fully Paid." This is the dependent (outcome) variable we are examining. Later in the analysis, we will transform this variable into "Loan Paid," which will then serve as our primary dependent variable.

Step 3: Identify Potential Independent (or Explanatory) Variables

In determining the best data to use, we must first consider the various explanatory, or independent, variables. That is, which variables are most likely to explain whether the loan was repaid or not? In this case, you may want to examine how much debt the borrower already has or how much income they have available to pay back the loan. For example, loan repayment might be determined by whether a borrower owns or rents the home, or still has a mortgage as specified in Column M, "home_ownership." Perhaps length of employment ("emp_length," Column L) and/or amount of annual income ("annual_inc," Column N) also help determine repayment. Maybe the debt-to-income ratio ("dti," Column Y) determines if a borrower will be able to repay the loan. Other possible independent variables include length of loan ("term," Column F), whether the income is verified ("verification_status," Column O), or state of residence ("addr_state," Column X). As the data analyst on this project, it is your job to identify independent variables that may help predict loan repayment.

Refer to the Lending Club data dictionary provided in Connect to determine which additional variables you think will affect loan repayment. (Note that not all variables listed in the data dictionary are included in this data set.)

Deliverable 1:

Once you have identified the variables that may explain loan repayment, prepare a document (in Microsoft Word) and save it as "Project 2: Lending Club Loan Analysis First Name and Last Name" (inserting your first and last name). Title the first section of the document "Dependent and Independent Variables." Next, within the document, identify and type the names of the dependent (outcome) variable and at least five proposed independent variables that you think explain loan repayment.

Step 4: Clean and Transform the Data in Preparation for Analysis

To prepare the data for analysis, proceed as follows:

1. *Make sure that data are not missing.* Check the data set to make sure that all variables desirable in the analysis are complete and are not missing data. (*Hint:* Run the COUNTA() function in Excel to see if the count of each variable equals the count of loans in the data set.)
2. *Make sure there are no errors in the data.* Using the Excel Data Analysis ToolPak (**Excel: Data > Analysis > Data Analysis > Descriptive Statistics**), run a descriptive analysis for your proposed independent and dependent variables looking at the minimums, maximums, and averages, or generate a histogram to identify unusual outliers or anomalies (see Chapter 4). Use the findings to assess whether the data contain any errors.

Deliverable 2:

In the second section of your report (after Deliverable 1), prepare a descriptive analysis titled "Descriptives" that includes a screenshot of descriptive statistics for all the proposed dependent and independent variables. Using the Data Analysis ToolPak, predict how the independent variables might relate to loan repayment. For example, we would predict a negative relationship between "dti" (the debt-to-income ratio) and loan repayment because the more debt you have, the less likely you'll be able to repay a loan.

3. *Transform the non-numeric data into numeric data.* Unless you are using exclusively categorical data, transform the data into numeric variables. Here are four examples of how to transform the variables:

   Loan Status: This is the variable we're using as our dependent (outcome) variable, but it appears as non-numeric data in the data set. Let's create a new variable titled "Loan Paid" with the value of 1 if the "loan_status" (Column Q) is "Fully Paid" and a value of 0 if the "loan_status" is "Charged Off." After this point in the analysis, we will use "Loan Paid" as the dependent variable, instead of "loan_status."

   Employment Length: The "emp_length" in Column L might be useful in the analysis as a number (number of years of employment), but in the data set it is non-numeric, with values including "3 years" and "10+ years." So you could transform the "10+ years" variable to have a numeric value of 10, transform "< 1 year" to have a numeric value of 0.5, and then name the new variable "Employment Length."

   Term of Loan: Presently, the term (Column F) has non-numeric values of "36 months" or "60 months." To transform the data into a new numeric variable (with values of either 30 or 60), reference the cell with the =LEFT() Excel command and make the first two digits of the "term" variable. Transform the term, and name the new variable "Term Length." Ensure that it is a numeric variable so that you can use it in your analysis.

   Verified: There are three possible values for Verification Status ("verification_status," Column O): "Verified," "Source Verified," or "Not Verified." Name the new transformed numeric variable "Verified," and transform the value equal to "1" if verification_status is "Verified" or "Source Verified" and a numeric value of "0" if "Not Verified."

Other transformations might also be useful for other independent variables. Consider each potential independent variable and transform it into a numeric variable that you think might be useful to analyze.

Deliverable 3:

In the third section of your report (after Deliverable 2), prepare a descriptive analysis titled "Transformation" that includes a description of any additional variables you transformed from non-numeric to numeric.

Analyze the Data

At this point, we have the data ready to predict loan repayment. Deciding which analytics techniques we will perform is critical to evaluating the relationship between the dependent (outcome) variable and the independent variables and ultimately to answering the question.

Step 1: Hypothesize Links Between Dependent/Outcome Variable (Loan Repayment) and the Independent Variables You Chose

What do you hypothesize are the links between the dependent (outcome) variable of loan repayment and the independent variables that might predict it? Look through the list of variables in the data dictionary in Connect. Which ones do you believe will predict loan repayment, and are they positively or negatively related to price?

Deliverable 4:

In the fourth section of your report (after Deliverable 3), predict how the independent variables might relate to loan repayment. For example, are people who own their home more likely or less likely to pay off a loan than those who rent? Make a prediction for each independent variable you think might be important and include them in your report.

Step 2: Select and Run Desired Analytics Techniques to Evaluate Relationships

What kind of analysis will best evaluate the relationship between our transformed Loan Paid (loan_status) and the various independent variables? Consider the following types of analyses:

- *Pivot tables:* Create a pivot table of Loan Paid ("loan_status" as dependent variable) and each independent variable you proposed (**Excel: Insert > Pivot Table**). After running the pivot analysis, do you still believe there is a relationship between the dependent variable and each independent variable you proposed?
- *Correlations:* Using the Data Analysis ToolPak (**Excel: Data > Analysis > Data Analysis > Correlation**), run a correlation between Loan Paid (dependent variable) and a few independent (now transformed) variables one by one. After running each correlation analysis, do you still believe there is a relationship between the dependent variable and each independent variable you proposed?
- *Regression analysis:* Using the Data Analysis ToolPak (**Excel: Data > Analysis > Data Analysis > Regression**), run a regression of Loan Paid (dependent variable) on your proposed independent (now transformed) variables. (For an additional description of regression analysis, see the statistics review in Chapter 3.) Recall that the independent variables must be grouped together if you perform this analysis in Excel.

Deliverable 5:

In the fifth section of your report after Deliverable 4, prepare an analysis titled "Analysis" that includes the following:

1. Screenshots of least two pivot table analyses and your interpretation (at least 100 words).
2. Screenshots of the output of at least two correlations between the dependent variable (Loan Paid) and one independent variable. What is your interpretation of the findings?
3. Screenshot of the results (output) of the regression, including the dependent variable (Loan Paid) and all of your proposed independent variables.
   a. What is your interpretation of the findings? (*Hint:* In general, if the *t*-statistic is greater than 2 or the *p*-value is less than 0.05, then the independent variable is likely a significant variable in predicting loan repayment.)
   b. Was the relationship in the direction (positive vs. negative vs. no relationship) different from what you anticipated in Deliverable 4? Explain.

Report the Results

Based on the analysis, consider how to present your findings explaining loan repayment to Lending Club management. Would you use graphs, reports, correlations, pivot tables, or regression results to report the results in the most effective way? Would you use tabulated data or visualizations to communicate the results? Explain your decisions.

Deliverable 6:

In the sixth section of your report after Deliverable 5, prepare an analysis titled "Final Report" that includes the following:

1. An executive summary that is a maximum of one page. It should concisely describe the recommended course of action and the main support for your recommendation.
2. A report summarizing your findings and any graphs, reports, correlations, pivot tables, or regression results that help communicate the findings effectively.
3. A brief discussion of how these findings might help predict loan repayment in the future even before a loan is extended.

12.3 PROJECT 3: COMPLETING YOUR OWN PROJECT USING THE SOAR ANALYTICS MODEL

LO 12.3

Propose your own business-related question and answer the question using the SOAR analytics model.

Now it is your turn to explore an area of interest to you using the SOAR analytics model. For this project, you ask the question, find appropriate data sources to address your question, analyze the data, and then report your findings to a decision-maker. You will take on the role of the analyst hired to provide recommendations to the management team of a business/organization on an accounting, finance, or other business-related issue. The question might relate to helping the company maximize profits, find the most profitable customers or most cost-effective worker, make the supply chain more efficient, market the product more effectively, and so on.

Using the SOAR Analytics Model

This is the scenario:

You have been asked to present an analysis asking and addressing an accounting, finance, marketing, operations, supply-chain, or other business-related question that you believe will help the company address an issue. Provide a write-up to the CEO of the company (or director of a nonprofit entity) containing a comprehensive summary of your analysis, as well as any recommendations for further action. Your presentation (at your instructor's option) should summarize your analysis, focusing on your recommendations and justification. At a minimum, your presentation and write-up should do the following.

1. *Provide background and an overview.* Summarize key facts about the business/organization. Describe key factors that affect the organization's performance as a profit-maximizing, governmental, or non-profit entity. If relevant, describe important current events affecting the organization that might give rise to your question.
2. *Explain your role and motivation.* What big-picture problem are you trying to address or solve with business analytics? Why is addressing this issue important to your organization?

3. *Specify the question.* Which specific questions are you trying to answer? How will you use business analytics to answer these questions? The more specific you can be with each question, the more specific your search for data and your data analysis can be. When asking a question, consider whether you can gain access to appropriate data that can answer that question.
4. *Obtain the data.* Which data set(s) will you use to answer the question? Consider using the data sources discussed in Chapter 2 and throughout the text, as well the other data sources listed in Exhibit 12.6. You will need to prepare the data: Clean it as necessary to prepare for analysis and make sure the data are loaded completely, aren't missing any information, and are machine readable, numeric, and ready for analysis. Are there any issues with data quality and other data manipulations, transformations, and additional attributes you developed?
5. *Analyze the data.* How will you analyze the data? Which type(s) of analytics are most applicable to your questions? Will you use descriptive, diagnostic, predictive, prescriptive, or adaptive analytics, or some combination of them? Which specific techniques will you use? What are the main limitations of your analyses? Are you concerned about data-quality issues in your data set?
6. *Report the results.* Provide a clear dashboard of key insights from your analysis. What are the key takeaways of your analyses? Provide two recommendations for the business/organization with supporting analyses/visualizations.
7. *Consider the future.* What additional actions or analyses would be useful to conduct in the future? What additional questions that are beyond the scope of your project would be useful to answer? If you could get any data possible to address these additional questions, what would those data sets include, and how would they help answer the questions?

You will be graded on (1) the *relevance* of your project in terms of identifying and addressing an important business/organizational/governmental issue, (2) the *novelty* of your approach, (3) the quality and depth of your *execution,* and (4) *effective* and *appropriate* use of the material learned in this text and course.

DELIVERABLES

The project deliverables are as follows:

Part 1: Report

Required (12-point font, single-spaced) as follows or as specified by your instructor:

1. *Executive summary.* Maximum of one page. This section should concisely describe the recommended course of action and the main support for your recommendation.
2. *Report.* Maximum of three pages. This section should detail the supporting analyses for your recommendations.
3. *Report appendix.* Maximum of three pages. This section should contain supporting tables, figures, and calculations for your report.

Part 2: Presentation

Required as follows, or as specified by your instructor:

1. A 10-minute presentation and 3- to 5-minute question-and-answer (Q&A) session.
2. All team members must participate in the presentation.

Potential Project Questions

To give you a sense of the types of questions you can ask, here are some potential project questions:

- Can I predict bankruptcy for internet companies based on Altman's Z? Can I predict bankruptcy for airlines based on their most recent financials?
- Can I forecast sales for **Facebook** or **Tesla** (using time series analysis)?
- Are time series projections of earnings more accurate than forecasts? Could I combine both into an earnings projection or market valuation?
- Can I predict monthly apartment rental prices based on location in the United States (or another country)?
- Are **Amazon** product reviews related to product sales?
- Is risk of loan repayment related to interest rate charged? (Use **Lending Club** data or other data.)
- What factors are associated with elderly workers staying in the workforce?
- Which is the more effective way to advertise: print media or social media?
- Which variables predict real-estate prices in national or local markets?
- What variables predict a company's stock market price, earnings, and/or operating cash flows?
- Is **Amazon** outperforming **Walmart** and the industry in terms of sales, earnings, stock returns, differences in debt, profit margins, and/or asset turnover?
- Which restaurant chain has been most affected by COVID-19?
- Which airline company has been most affected by COVID-19? (Evaluate using available seat miles or sales overall.)
- Which entertainment company has been most affected by COVID-19 in terms of sales, profitability, and stock returns?

Category	Data Set Description	Website
Economics	BRICS World Bank Indicators (Brazil, Russia, India, China, and South Africa)	https://www.kaggle.com/docstein/brics-world-bank-indicators
Economics	Bureau of Economic Analysis Data	http://www.bls.gov/data/
Economics	Commodities Trading Data	https://www.macrotrends.net/charts/commodities
Economics	Exchange Rates	https://www.macrotrends.net/charts/exchange-rates
Economics	Macroeconomy (Housing Starts, Unemployment, Auto Sales, etc.)	https://www.macrotrends.net/charts/economy
Financial	Financial Statement Data	https://www.calcbench.com/
Financial	Financial Statement Data, EDGAR, Securities and Exchange Commission	https://www.sec.gov/edgar.shtml
Financial	Analyst Forecasts	Yahoo! Finance (finance.yahoo.com), Analysis Tab
Financial	Stock Market Data Set	https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs
Financial	Credit Card Fraud Detection	https://www.kaggle.com/mlg-ulb/creditcardfraud
Financial	Daily News/Stock Market Prediction	https://www.kaggle.com/aaron7sun/stocknews
Financial	Stock Market Indices	https://www.macrotrends.net/charts/stock-indexes
Financial	Interest Rates	https://www.macrotrends.net/charts/interest-rates

Exhibit 12.6
Potential Sources of Data to Address Business Questions

Category	Data Set Description	Website
Financial	Retail Data Analytics	https://www.kaggle.com/manjeetsingh/retaildataset
Financial	Peer-to-Peer Lending Data of Approved and Rejected Loans	lendingclub.com (requires login)
Financial	Daily Stock Prices (and Weekly and Monthly)	Yahoo! Finance (finance.yahoo.com), Historical Data Tab
Financial	Financial and Economic Summaries by Industry	http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datacurrent.html
Financial Aid	Cohort Default Rates by School, Lender, State, and Institution Type	https://studentaid.gov/data-center/student/default
General	data.world	https://data.world/
General	kaggle.com	https://www.kaggle.com/datasets
Government	State of Ohio Financial Data	http://https://checkbook.ohio.gov/
Government	City of Chicago Financial Data	https://data.cityofchicago.org
Government	City of New York Financial Data	http://www.checkbooknyc.com/spending_landing/yeartype/B/year/119
Marketing	Amazon Product Reviews	https://data.world/datafiniti/consumer-reviews-of-amazon-products
Other	Restaurant Safety	https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j
Other	Citywide Payroll Data	https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e
Other	Property Valuation/Assessment	https://data.cityofnewyork.us/City-Government/Property-Valuation-and-Assessment-Data/yjxr-fw8i
Other	Country-by-Country Statistics	https://www.irs.gov/statistics/soi-tax-stats-country-by-country-report
Other	Interesting Fun Data Sets—14 Data Science Projects with Data	https://towardsdatascience.com/14-data-science-projects-to-do-during-your-14-day-quarantine-8bd60d1e55e1
Other	Links to Big Data Sets—Amazon Web Services	https://aws.amazon.com/public-datasets/
Population	Largest Countries by Population	https://www.macrotrends.net/countries/ranking/population
Real Estate	New York Airbnb Data Explanation	https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data
Real Estate	U.S. Airbnb Data	https://www.kaggle.com/kritikseth/us-airbnb-open-data/tasks?taskId=2542
Real Estate	Trip Advisor Hotel Reviews	https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews
Real Estate	Monthly Real Estate Sales Data by Metro Area	https://www.realtor.com/research/data/
Retail	Retail Sales Forecasting	https://www.kaggle.com/tevecsystems/retail-sales-forecasting

Exhibit 12.6 (Concluded)

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

APPENDIX A

Excel Tutorial (Formatting, Sorting, Filtering, and Pivot Tables)

You may already have a basic understanding of Excel from your other classes. As a review, and to gain a few more skills, we offer this basic tutorial, which has the following elements:

1. Basic formatting for an income statement using Excel function SUM().
2. Basic data manipulation (filters, sorts, pivot tables).

Basic Formatting for an Income Statement Using Excel Function SUM()

A company's income statement summarizes its sales (revenues), expenses, and profit or loss for a particular period, such as a quarter or a year. Suppose we want to put the following data into the appropriate income-statement format:

Revenues	50,000
Expenses	
Cost of Goods Sold	20,000
Research and Development Expenses	10,000
Selling, General, and Administrative Expenses	10,000
Interest Expense	3,000

Required

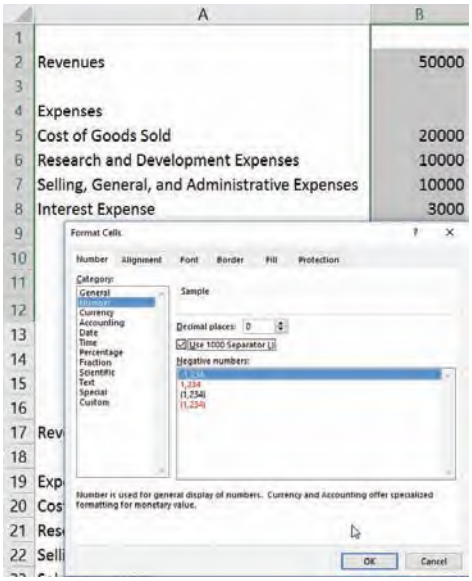
1. Add a comma as a 1000 separator for each number.
2. Insert the words "Total Expenses" below the list of expenses.
3. Calculate subtotal for Total Expenses using the **SUM()** command.
4. Insert a single bottom border under Interest Expense and under the Total Expenses subtotal.
5. Insert the words "Net Income" and calculate Net Income (Revenues – Total Expenses).
6. Format the top and bottom numbers of the column with a $ currency sign.
7. Insert a bottom double border to underline the final Net Income total.

Solution

1. Open Appendix_A_Data.xlsx in Excel. Highlight the column with all the numbers. Right-click on Format Cells. . . to open the dialog box shown in Exhibit A1.

   Click on Number and set decimal places to zero. Click on Use 1000 Separator (,) and click OK.

Exhibit A1



2. Insert the words "Total Expenses" below the list of expenses.
   Type "Total Expenses" at the bottom of the list of expenses.
3. Calculate subtotal for Total Expenses using the SUM() command.

Use the SUM() command to sum all the expenses, as shown in Exhibit A2. The resulting output appears in Exhibit A3.

Exhibit A2

	A	B
1		
2	Revenues	50,000
3		
4	Expenses	
5	Cost of Goods Sold	20,000
6	Research and Development Expenses	10,000
7	Selling, General, and Administrative Expenses	10,000
8	Interest Expense	3,000
9	Total Expenses	=SUM(B5:B8)

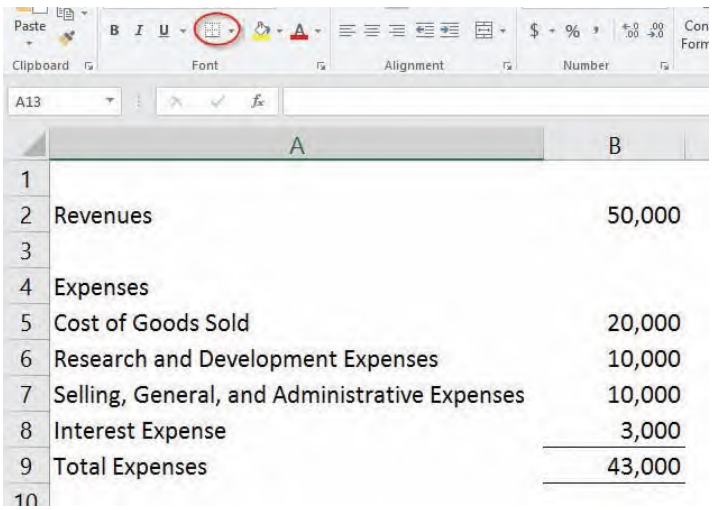
Exhibit A3

Revenues	50,000
Expenses	
Cost of Goods Sold	20,000
Research and Development Expenses	10,000
Selling, General, and Administrative Expenses	10,000
Interest Expense	3,000
Total Expenses	43,000

4. Insert a single bottom border under Interest Expense and under the Total Expenses subtotal.

Use the border icon located in the Font toolbar on the Home tab to add the bottom border (as shown in Exhibit A4).

Exhibit A4



	A	B
1		
2	Revenues	50,000
3		
4	Expenses	
5	Cost of Goods Sold	20,000
6	Research and Development Expenses	10,000
7	Selling, General, and Administrative Expenses	10,000
8	Interest Expense	3,000
9	Total Expenses	43,000
10		

5. Insert the words "Net Income" and calculate Net Income (Revenues – Total Expenses).

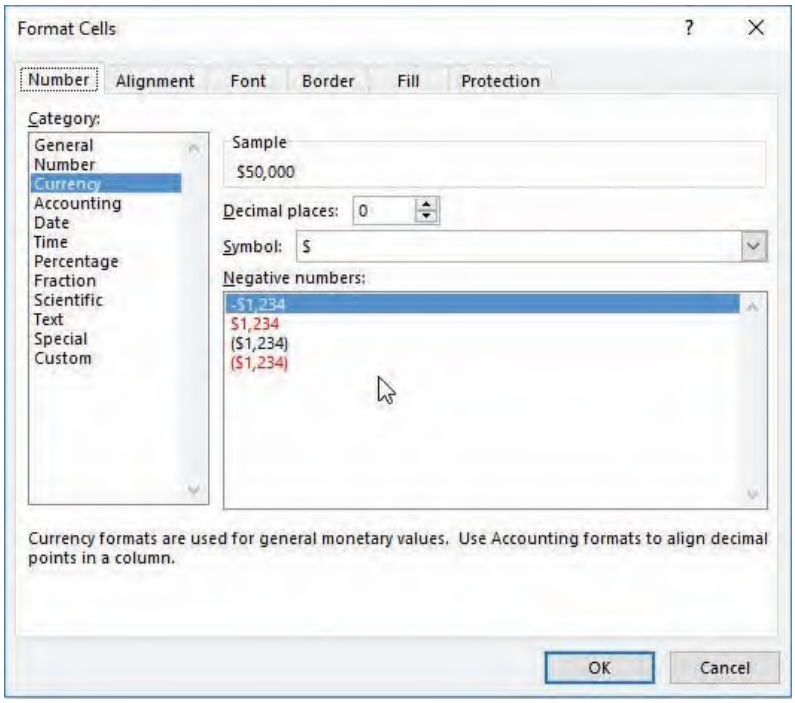
Type "Net Income" at the bottom of the spreadsheet. Calculate net income by inserting the correct formula in the cell (here, "=B2-B9"), as shown in Exhibit A5.

Exhibit A5

	A	B
1		
2	Revenues	50,000
3		
4	Expenses	
5	Cost of Goods Sold	20,000
6	Research and Development Expenses	10,000
7	Selling, General, and Administrative Expenses	10,000
8	Interest Expense	3,000
9	Total Expenses	43,000
10		
11	Net Income	=B2-B9

6. Format the top and bottom numbers of the column with a $ currency sign. Right-click on each number and Format Cells, select Currency and no decimal places (as shown in Exhibit A6).

Exhibit A6



7. Insert a bottom double border to underline the final Net Income total. Place your cursor on the cell containing Net Income (7,000). Then select Bottom Double Border from the **Font > Borders** menu.

The final project is shown in Exhibit A7.

Exhibit A7

Revenues	$50,000
Expenses	
Cost of Goods Sold	20,000
Research and Development Expenses	10,000
Selling, General, and Administrative Expenses	10,000
Interest Expense	3,000
Total Expenses	43,000
Net Income	$7,000

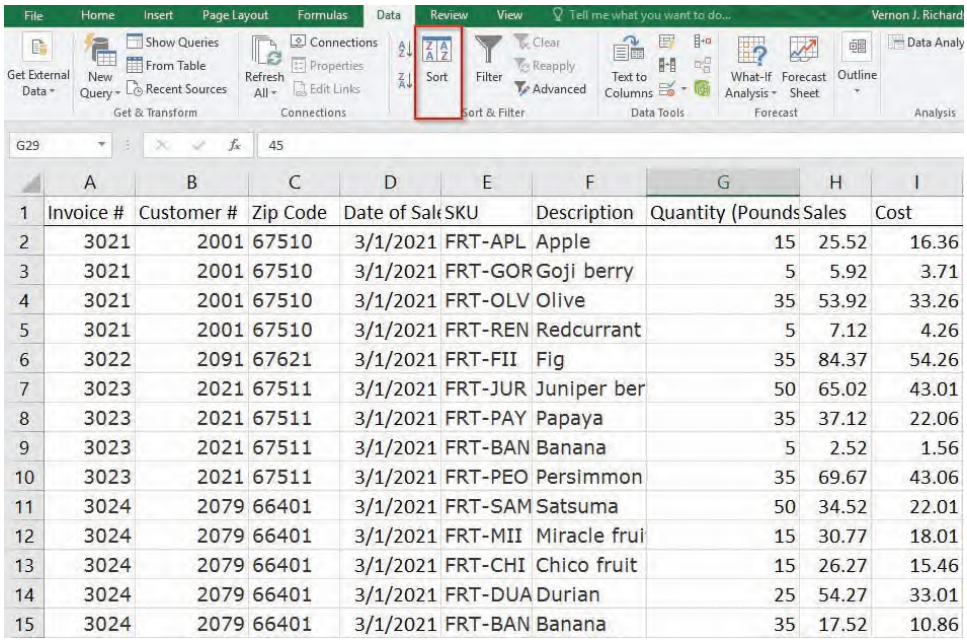
BASIC DATA MANIPULATION (SORTS, FILTERS, AND PIVOT TABLES)

1. Open Lab 0 Appendix Data.xlsx.
2. Look at the data.

Sorting the Data

3. Sort the data. To do so, select **Data > Sort & Filter > Sort** (as shown in Exhibit A8).

Exhibit A8



	A	B	C	D	E	F	G	H	I
1	Invoice #	Customer #	Zip Code	Date of Sale	SKU	Description	Quantity (Pounds	Sales	Cost
2	3021	2001	67510	3/1/2021	FRT-APL	Apple	15	25.52	16.36
3	3021	2001	67510	3/1/2021	FRT-GOR	Goji berry	5	5.92	3.71
4	3021	2001	67510	3/1/2021	FRT-OLV	Olive	35	53.92	33.26
5	3021	2001	67510	3/1/2021	FRT-REN	Redcurrant	5	7.12	4.26
6	3022	2091	67621	3/1/2021	FRT-FII	Fig	35	84.37	54.26
7	3023	2021	67511	3/1/2021	FRT-JUR	Juniper ber	50	65.02	43.01
8	3023	2021	67511	3/1/2021	FRT-PAY	Papaya	35	37.12	22.06
9	3023	2021	67511	3/1/2021	FRT-BAN	Banana	5	2.52	1.56
10	3023	2021	67511	3/1/2021	FRT-PEO	Persimmon	35	69.67	43.06
11	3024	2079	66401	3/1/2021	FRT-SAM	Satsuma	50	34.52	22.01
12	3024	2079	66401	3/1/2021	FRT-MII	Miracle frui	15	30.77	18.01
13	3024	2079	66401	3/1/2021	FRT-CHI	Chico fruit	15	26.27	15.46
14	3024	2079	66401	3/1/2021	FRT-DUA	Durian	25	54.27	33.01
15	3024	2079	66401	3/1/2021	FRT-BAN	Banana	35	17.52	10.86

4. Sort by sales price from largest to smallest. Input Sales into the Sort by, select Largest to Smallest in the dialog box, and click OK (as shown in Exhibit A9). The highest sales price appears to be apricots at a cost of $140 (Exhibit A10). Looking down at the bottom of this list, we see that the lowest sales price appears to be bananas for $2.52 (not shown in Exhibit A10).

Exhibit A9

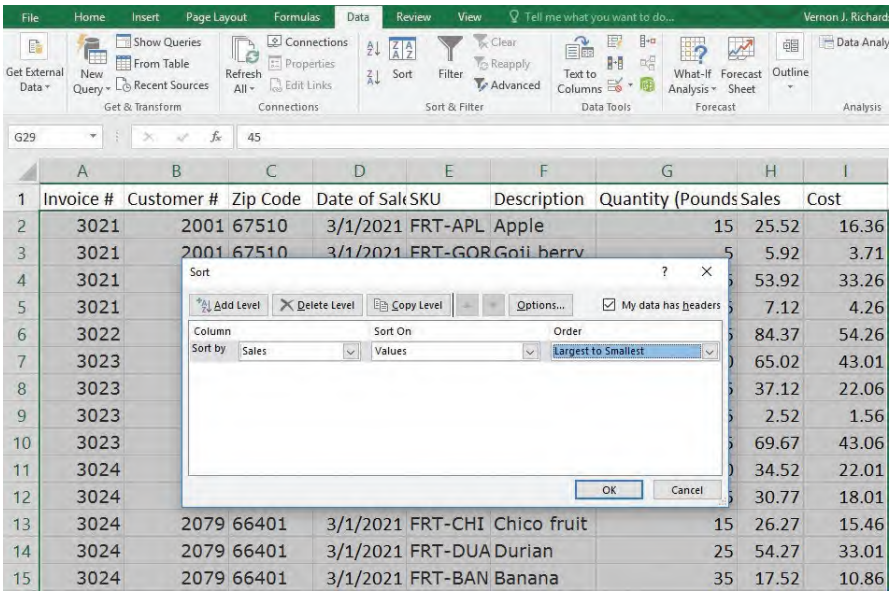


Exhibit A10

	A	B	C	D	E	F	G	H	I
1	Invoice #	Customer #	Zip Code	Date of Sale	SKU	Description	Quantity (Pounds	Sales	Cost
2	3069	2047	67838	3/4/2021	FRT-APO	Apricot	50	140	88.01
3	3026	2088	67004	3/1/2021	FRT-FII	Fig	50	120.5	77.51
4	3043	2087	67731	3/2/2021	FRT-TAN	Tamarind	50	109.5	69.01
5	3028	2086	67416	3/1/2021	FRT-COU	Coconut	45	108.5	69.31
6	3046	2064	67732	3/2/2021	FRT-SOO	Soursop	45	108	65.71
7	3039	2060	67518	3/2/2021	FRT-PLT	Plumcot (or	50	105	62.01
8	3067	2072	66736	3/4/2021	FRT-LYE	Lychee	50	100.5	62.51
9	3069	2047	67838	3/4/2021	FRT-BIT	Bilberty	50	100	58.51
10	3068	2046	67035	3/4/2021	FRT-LEO	Lemon	40	98.42	59.21
11	3048	2031	67626	3/2/2021	FRT-SOO	Soursop	40	96.02	58.41

Filtering the Data

Next, filter the data to look at only the banana transactions.

5. To sort the data, select **Data > Sort & Filter > Filter.**
6. An upside-down triangle (called a *chevron*) will appear. Click the chevron in cell F1, click Select All to unselect all, and then select only the word "Banana." The resulting data should appear as shown in Exhibit A11.

Exhibit A11

Invoice #	Customer #	Zip Code	Date of S	SKU	Description	Quantity (Poun	Sales	Cost
3041	2066	66408	3/2/2021	FRT-BAN	Banana	50	25.02	15.51
3025	2099	66710	3/1/2021	FRT-BAN	Banana	45	22.52	13.96
3055	2081	66014	3/3/2021	FRT-BAN	Banana	45	22.52	13.96
3028	2086	67416	3/1/2021	FRT-BAN	Banana	40	20.02	12.41
3054	2088	66842	3/2/2021	FRT-BAN	Banana	40	20.02	12.41
3069	2047	67838	3/4/2021	FRT-BAN	Banana	40	20.02	12.41
3024	2079	66401	3/1/2021	FRT-BAN	Banana	35	17.52	10.86
3048	2031	67626	3/2/2021	FRT-BAN	Banana	35	17.52	10.86
3049	2043	66840	3/2/2021	FRT-BAN	Banana	25	12.52	7.76
3048	2031	67626	3/2/2021	FRT-BAN	Banana	15	7.52	4.66
3050	2084	66936	3/2/2021	FRT-BAN	Banana	15	7.52	4.66
3061	2024	66937	3/3/2021	FRT-BAN	Banana	15	7.52	4.66
3023	2021	67511	3/1/2021	FRT-BAN	Banana	5	2.52	1.56

7. Alternatively, we could filter based on date to get all transactions on 3/2/2021. We first need to clear the filter in cell F1, by clicking on the Filter symbol and selecting "Select All."
8. Click the chevron in cell D1, click Select All to unselect all, and then select "2021," then "March," then "2". The sorted results are as shown in Exhibit A12.

Exhibit A12

	A	B	C	D	E	F	G	H	I
1	Invoice #	Customer #	Zip Code	Date of S	SKU	Description	Quantity (Poun	Sales	Cost
4	3043	2087	67731	3/2/2021	FRT-TAN	Tamarind	50	109.5	69.01
6	3046	2064	67732	3/2/2021	FRT-SOO	Soursop	45	108	65.71
7	3039	2060	67518	3/2/2021	FRT-PLT	Plumcot (or	50	105	62.01
11	3048	2031	67626	3/2/2021	FRT-SOO	Soursop	40	96.02	58.41
12	3045	2100	66226	3/2/2021	FRT-PEA	Pear	50	91.52	53.51
14	3043	2087	67731	3/2/2021	FRT-CUE	Cucumber	50	91.02	54.51
15	3053	2032	66414	3/2/2021	FRT-JAU	Jambul	45	87.77	50.86
16	3043	2087	67731	3/2/2021	FRT-APL	Apple	50	85.02	54.51
26	3046	2064	67732	3/2/2021	FRT-FEO	Feijoa	40	79.22	52.01

Pivot Tables

Gross margin: Net sales less the cost of goods sold. It is the amount of money a company retains after deducting the direct costs of producing a good or service.

Let's compute the accumulated gross margin for bananas, apricots, and apples. We will take a few steps to get the data ready for pivot tables.

9. First, unclick the filter at **Data > Sort & Filter > Filter** by clicking on and unselecting Filter.
10. Next, compute the gross margin for each line item in the invoice. In cell J1, input the words "Gross Margin". Underline it with a bottom border. In cell J2, input the formula "=H2-I2" and hit Enter. See Exhibit A13.

	A	B	C	D	E	F	G	H	I	J
1	Invoice #	Customer #	Zip Code	Date of Sale	SKU	Description	Quantity (Pounds	Sales	Cost	Gross Margin
2	3069	2047	67838	3/4/2021	FRT-APO	Apricot	50	140	88.01	=H2-I2
3	3026	2088	67004	3/1/2021	FRT-FII	Fig	50	120.5	77.51	
4	3043	2087	67731	3/2/2021	FRT-TAN	Tamarind	50	109.5	69.01	

Exhibit A13

11. Copy the result from cell J2 to cells J3:J194.
    Now it is time to use a pivot table, which summarizes selected columns in a spreadsheet but doesn't change the spreadsheet itself. Recall we are trying to summarize the accumulated gross margin for bananas, apricots, and apples.
12. Select **Insert > Tables > PivotTable.**

Analytics Tool: Excel Pivot Tables

PivotTables allow you to quickly summarize large amounts of data. In Excel, select **Insert** > **PivotTable,** choose your data source, and then click the check mark next to or drag your fields to the appropriate boxes in the PivotTable Fields pane to identify filters, columns, rows, or values. You can easily move attributes from one pane to another to quickly "pivot" your data. Here is a brief description of each section.

Rows: Show the main item of interest. You usually want master data here, such as customers, products, or accounts.

Columns: Slice the data into categories or buckets. Most commonly, columns are used for time (for example, years, quarters, months, dates).

Values: This area represents the meat of your data. Any measure that you would like to count, sum, average, or otherwise aggregate should be placed here. The aggregated values will combine all records that match a given row and column.

Filters: Placing a field in the Filters area will allow you to filter the data based on that field, but it will not show that field in the data. For example, if you want to filter based on a date, but you don't care to view a particular date, you could use this area of the field list. More recent versions of Excel offer improved methods for filtering, but this legacy feature is still functional.

13. Make sure all data are selected as follows in Table/Range and click OK (as shown in Exhibit A14).

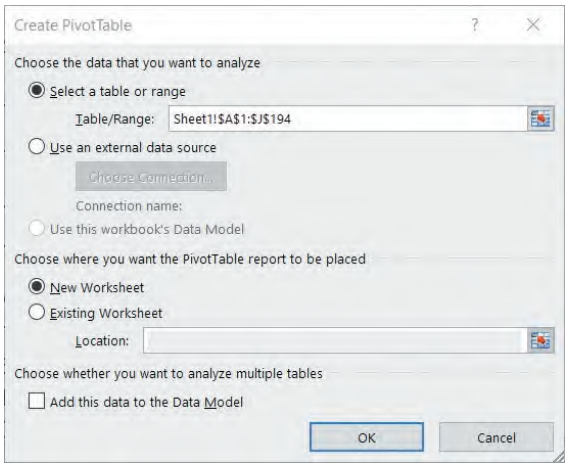


Exhibit A14

14. The empty pivot table will open in a new worksheet, ready for the pivot table analysis. To build a report, choose fields from the pivot table field list.

Inputs into Pivot Table

Columns:
Rows: [Description]
ΣValues: [Gross Margin]

Drag [Description] from PivotTable Fields into the Rows and [Gross Margin] from FIELD NAME into ΣValues fields in the pivot table. The ΣValues will default to "Sum of Gross Margin."

The resulting pivot table will look as shown in Exhibit A15. The analytics suggests that the gross margin for apples is $140.39; for apricots, $78.02; and for bananas, $77.08.

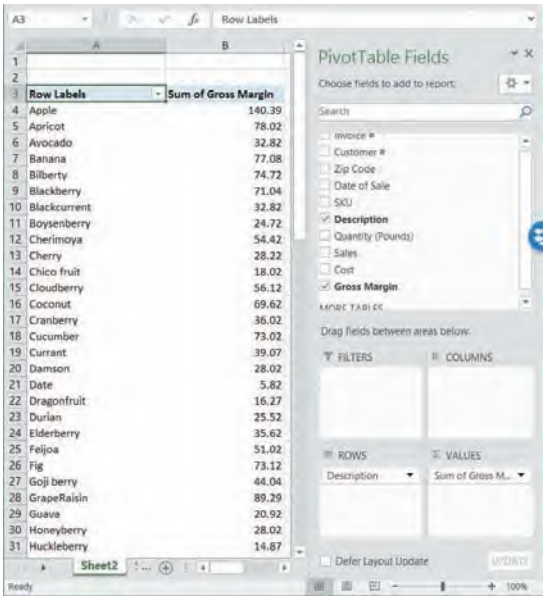


Exhibit A15

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

APPENDIX B

Tableau Tutorial

Tableau is a tool for data visualization. Using its drag-and-drop functionality to showcase your data is similar to using a pivot table, but its default is to visualize your data, not to show it in a tabular format the way a pivot table does.

Exhibit B1 provides a quick introduction to pieces of the Tableau canvas.

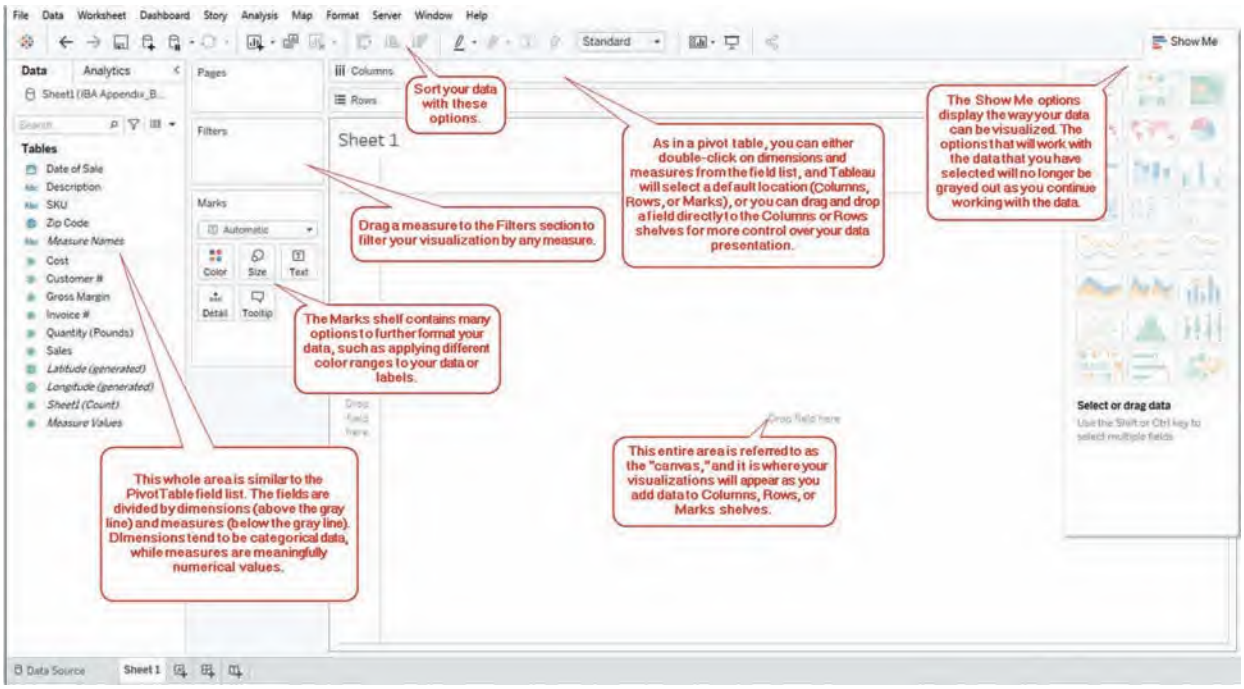


Exhibit B1

Tableau Software

Tableau can connect to a variety of data types, including data housed in Excel, Access, and SQL Server. We will connect to the data set Appendix_B_Data.xlsx.

1. Open Tableau.
2. Immediately upon opening Tableau, you will see a list of file types that you can connect to. We'll connect to an Excel file, so click Microsoft Excel (as shown in Exhibit B2).

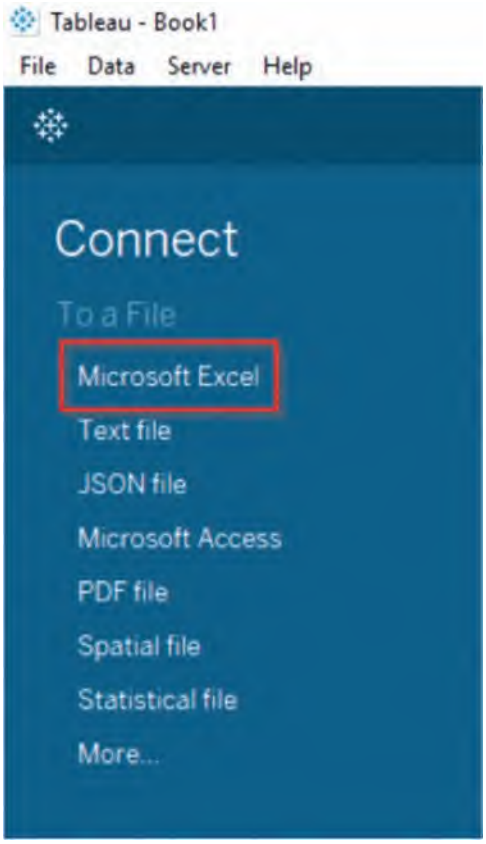


Exhibit B2

Tableau Software

3. Navigate to where your file is stored, highlight the file, and click Open.

Tableau automatically detects the data types of the attributes you import. In this data set, the attributes probably all imported as the data type you would expect. Notice that the first two, Invoice # and Customer #, imported as numbers. Continue looking at the attributes, and you will notice the globe icon above Zip Code (Exhibit B3). This is Tableau showing you one of its best features: The Zip Code data were imported as geographic data, which will allow you to create maps.

# Sheet1 Invoice #	# Sheet1 Customer #	Sheet1 Zip Code	Sheet1 Date of Sale	Abc Sheet1 SKU	Abc Sheet1 Description
3,069	2,047	67838	3/4/2021	FRT-APO	Apricot
3,026	2,088	67004	3/1/2021	FRT-FIl	Fig
3,043	2,087	67731	3/2/2021	FRT-TAN	Tamarind

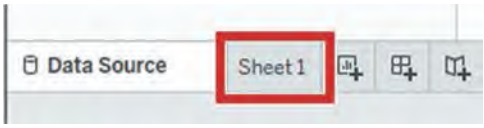
Exhibit B3

Tableau Software

4. To begin working with the data, click Sheet 1 in the bottom left (as shown in Exhibit B4).

Exhibit B4

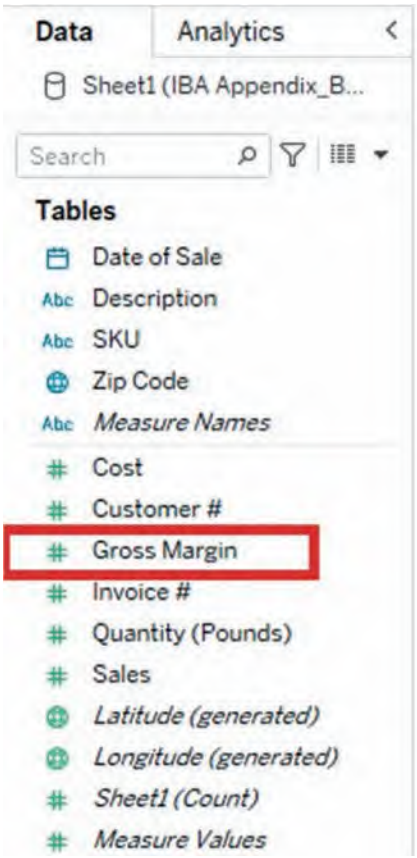
Tableau Software



5. Now double-click on the measure Gross Margin (as shown Exhibit B5).

Exhibit B5

Tableau Software



Immediately you will see that Tableau interacts with data differently than Excel because it has defaulted to displaying a bar chart. This isn't a very meaningful chart as it is, but you can add meaning by adding a dimension.

6. Double-click Description from the dimensions. You find the same numbers that you find in the analytics in Appendix A: apples with a gross margin of 140.4, apricots with a gross margin of 78.0, and so on.
7. To make these data easier to interpret, you can sort them. Click the Sort Descending icon (Exhibit B6) to sort the data.



Exhibit B6

Tableau Software

You can continue adding attributes to the Columns or Rows shelves and changing the method of visualization by using the Show Me tab to further familiarize yourself with Tableau.

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

APPENDIX C

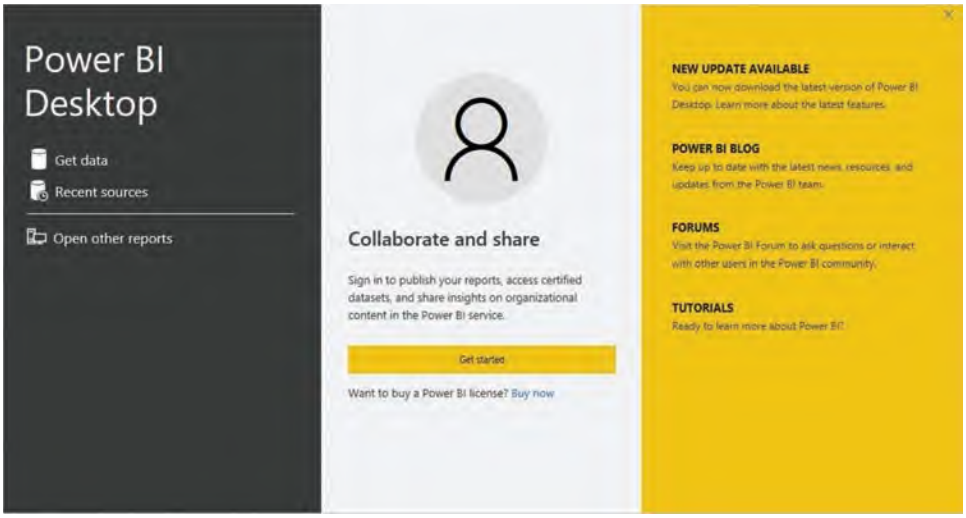
Power BI Desktop Tutorial

Power BI Desktop is a Microsoft tool that combines ETL (Extract, Transform, Load) tools with reporting tools. When we work with Power Query or PowerPivot in Excel, we're actually working with Power BI tools. If you run statistical tests such as hypothesis testing or regression analysis, it's best to work in Excel directly and use the Power Query add-in. However, if you need to transform your data using Power Query prior to creating reports or dashboards, or even if you just want to explore your data, Power BI Desktop is a great tool.

Power BI creates data visualizations that are very similar to the visualizations you can create in Tableau, but the path to getting there is different. Power BI defaults to a report mode (similar to Tableau's Dashboard mode), so that as you create visuals, they appear as tiles that you can resize and rearrange around the canvas.

When you open Power BI Desktop, you will see a start-up screen similar to the one shown in Exhibit C1.

Exhibit C1



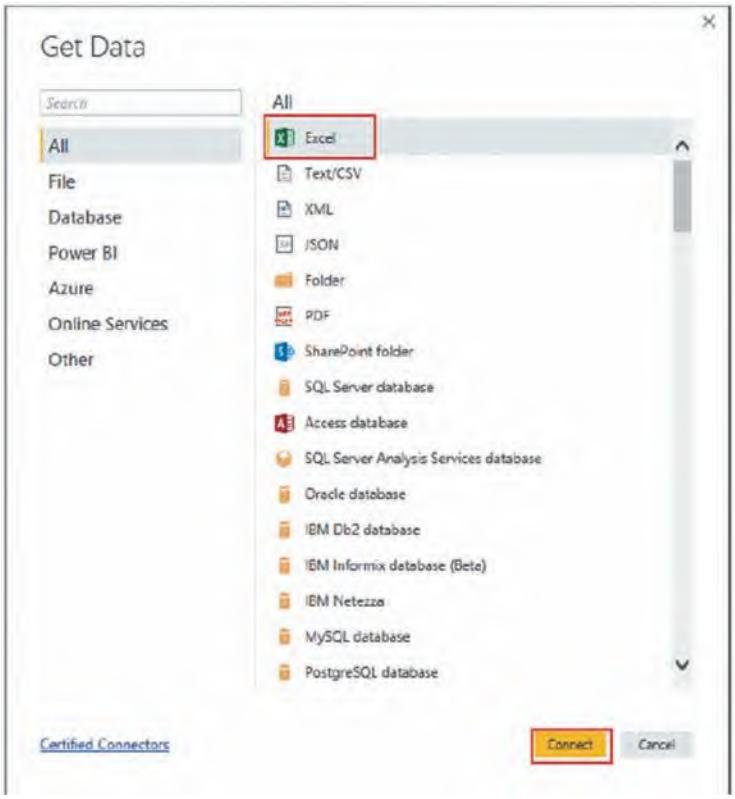
The tutorials and other training resources on the right-hand side of the start-up screen are helpful for getting started with Power BI.

The Get Data button on the left-hand side of the start-up screen will bring you into Power BI's Power Query tool. It is set up exactly like Excel's Power Query Tool, so you can use it to connect to a variety of sources, including Excel, SQL Server, and Access.

To familiarize you with Power BI, we will use the data in Appendix_C_Data.xlsx. The data are a subset of the sales data related to a fictitious brewery named Sláinte.

1. Click Get Data on the start-up screen.
2. Select Excel from the list of possible data sources, then click Connect (as shown in Exhibit C2).

Exhibit C2



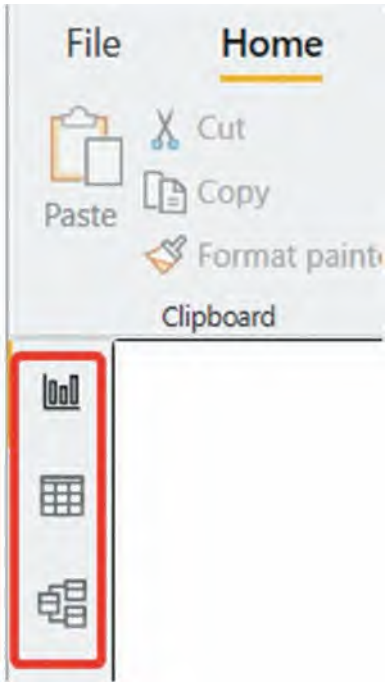
3. Browse to the file location for Appendix_C_Data.xlsx and open the file.
4. Because there are three spreadsheets in the file, the Navigator gives you the option to select 1, 2, or all of the spreadsheets. Place check marks next to each.
5. You are also given an option to either Load or Edit the data. If you click Edit, you will enter the Power Query window, which has the same ribbon and options to transform the data as the Excel version of the tool (add columns, split columns, pivot data, and so on). These data do not need to be transformed, so click Load.
6. Once the data are loaded, you will see a blank canvas on which you can build a report. There are three key elements bordering the canvas:
   a. To the left of the blank canvas, you are presented with the three options shown in Exhibit C3.

      Report Mode: The first option, represented by an icon that looks like a bar chart, is for Report mode. This is the default view, and it is where you can build visualizations and explore data.

      Data Mode: The second option, represented by an icon that looks like a table or a spreadsheet, is for Data mode. If you click on this icon, you can view the raw data that you have imported into Power BI. In this mode, you can also create new measures or new columns.

      Model Mode: The third option, which looks like a database diagram, is for Model mode. When you click on this icon, you enter PowerPivot. From this mode, you can edit the table, attribute names, and relationships between tables.

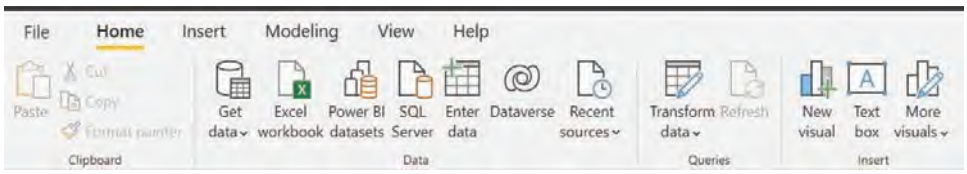
Exhibit C3



b. To the right of the blank canvas are your Visualizations and Fields list.
Visualizations: You can drag any of these options into the canvas to begin designing a visualization. Once you have tiles on your report, you can change the type of visualization used to depict a set of fields by clicking the tile and then selecting a different visualization option.
Fields: This section is similar to the PivotTable field list. You can expand the tables to see their attributes. Placing a check mark in the fields will add them to an active tile.
Values, Filters, etc.: This section varies based on the tile and the fields you are actively working with. Any time you add a field to a visualization, that field gets automatically added to the filters, which eliminates the need to manually add filters or slicers to your PivotTable.

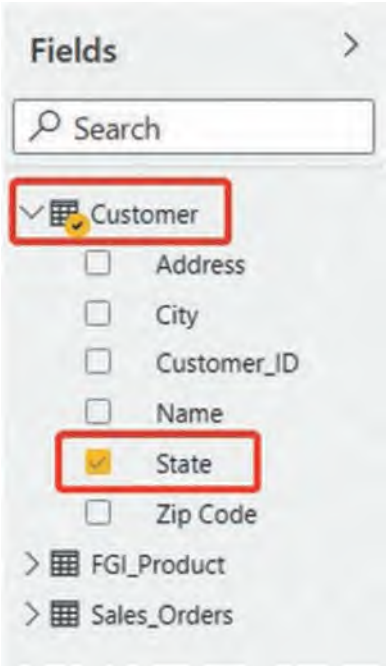
c. Immediately above the canvas is the familiar ribbon that you can expect from Microsoft applications. The available tabs differ across the three different modes (Report, Data, and Model), but four tabs—Home, Modeling, View, and Help—stay consistent across the three modes. Exhibit C4 displays the ribbon that is available in the Report mode.

Exhibit C4



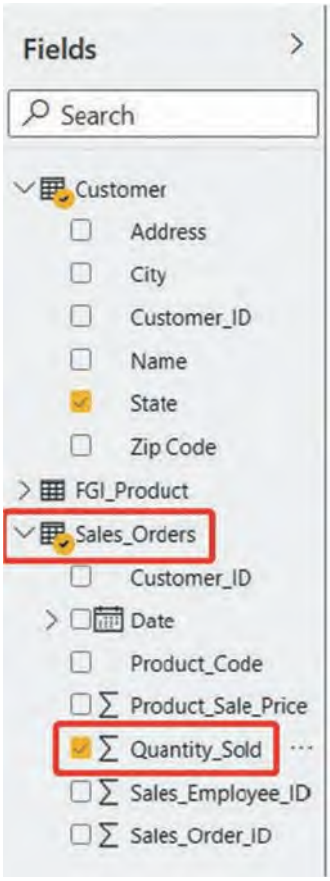
7. To begin working with the data, expand the Customer table and place a check mark in the State field (as shown in Exhibit C5).

Exhibit C5



8. Power BI will default to creating a tile with a map visualization, similar to how Tableau defaults to working with geographic data. To make the map more interesting, expand the Sales_Orders table and place a check mark in the Quantity_Sold field (as shown in Exhibit C6).

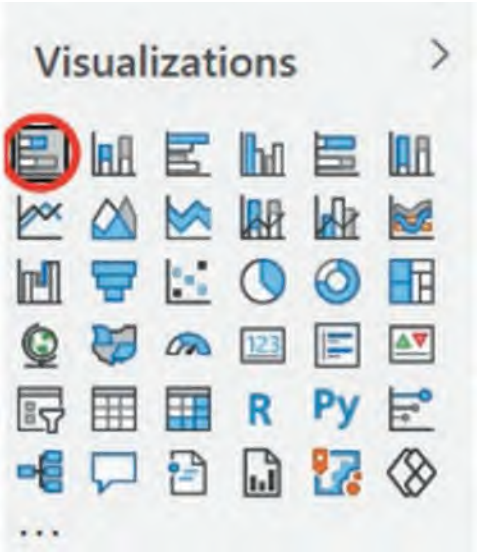
Exhibit C6



This will make the tile more interesting by changing the size of the symbol associated with each state—the larger the symbol, the higher the quantity sold in that state.

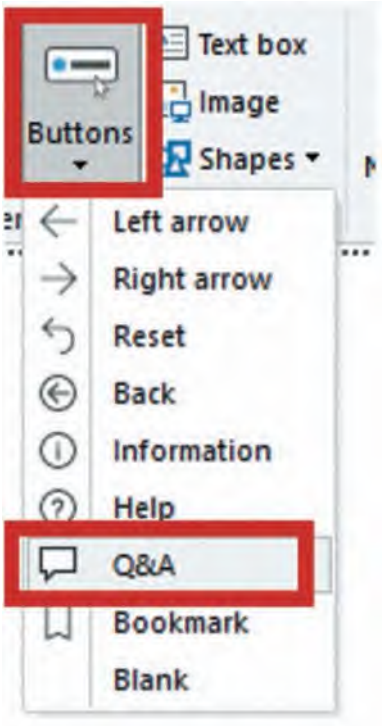
9. You can also change the way the data are presented by selecting a different visualization type. Select the first option in the Visualizations window (as shown in Exhibit C7) to view the data in a horizontal bar chart.

Exhibit C7



10. One of Power BI's most exciting offerings is its natural language processing for data exploration. Click the Insert tab in the ribbon, then click Buttons. In the drop-down, select Q&A (Exhibit C8).

Exhibit C8



11. The Q&A icon shown in Exhibit C9 will appear as a separate tile. If the placement defaults to being on top of the bar chart, you can click and drag it to somewhere else on the canvas.



Exhibit C9

12. To activate the Q&A, ctrl + click the icon. The window shown in Exhibit C10 will pop up. You can select from the list of questions that Power BI has generated, or you can type directly into the "Ask a question about your data" box.

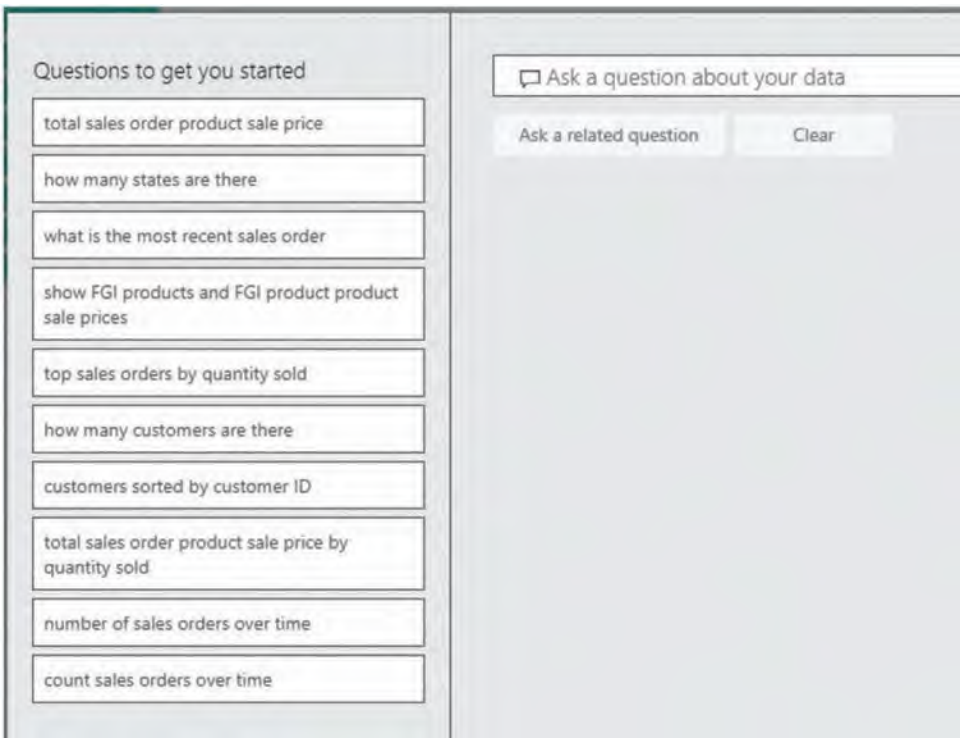
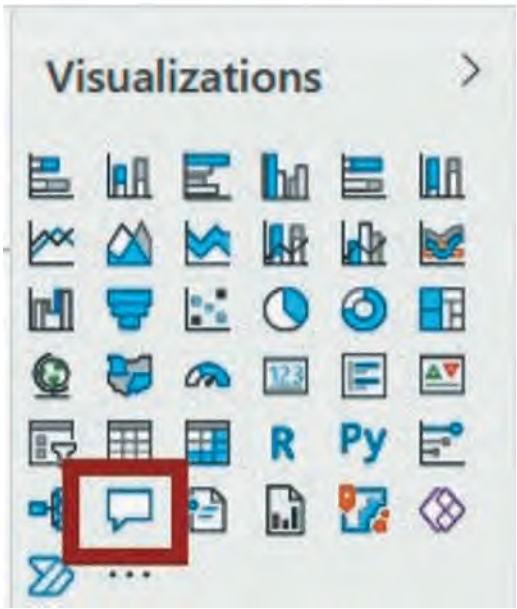


Exhibit C10

13. You can also add a question directly to the canvas. To do so, select the Q&A icon from the Visualizations pane (Exhibit C11).

Exhibit C11



There are many other exciting features in Power BI. This introduction should give you the confidence to jump in and explore what Power BI has to offer.

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Basic Statistics Tutorial

This appendix provides a review of the formulas that underlie the basic statistics that are used in this textbook.

Summary of Basic Statistics Terms

Some symbols used for different statistical terms vary based on whether they reference a sample statistic or a population parameter. Because we usually do not have data for a full population, we are more likely to work with sample data than with population data. For this reason, the formulas in this appendix are primarily based on sample data.

Description	Sample	Population
Size (number of observations)	n	N
Mean (or average)	$\bar{x}$	μ
Standard deviation	s	σ
Proportion	p	P
Sum	Σ	
Alpha (significance level)	α	

Measures of Central Tendency

To describe a sample, we often use measures of the central, middle, or most typical value. The mean, median, and mode are three common measures used to assess central tendency.

Term	Definition	Formula or Procedure
Mean (Average)	The sum of all values in a data set divided by the number of observations	$\frac{\Sigma x_i}{n}$
Median	Middle point in a set of observations	Rank order the observations from lowest to highest, and identify the observation in the middle. If there is an even number of observations, take the average of the middle two observations.
Mode	The most frequently occurring observation in a data set	Rank order the observations from lowest to highest and tally the number of times each observation occurs. The observation with the highest tally is the mode. A data set can be multimodal if there is more than one observation with the highest number of observations.

Measures of Dispersion

After describing the central tendency of the data, the next step is to assess its dispersion, spread, or variability. Three important measurements are the standard deviation, variance, and range.

Term	Definition	Formula
Standard Deviation	Indicates the difference between observations and the mean. The calculation is the square root of the variance. It is reported in the same unit of measure as the original data, making it easy to interpret.	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Variance	Indicates the difference between observations and the mean, defined as the sum of squared deviations from the mean or sum of the squares.	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Range	Maximum value in a data set minus the minimum value in the dataset	Maximum observation – minimum observation

Confidence Intervals[1]

Instead of basing an inference on a single value (such as a sample mean), a confidence interval provides a range of values, which (hopefully) contains the true population parameter at a certain level of confidence (for example, 90 percent, 95 percent). Confidence intervals are made up of two parts, the point estimate and the margin of error. The *point estimate* is the number we use from the sample to estimate the population parameter. The *margin of error* reflects random error associated with sampling and is the amount added to and subtracted from the point estimate to construct the confidence interval. Thus, the confidence interval equals:

$$\text{Confidence Interval} = \text{Point estimate} \pm \text{Margin of error}$$

The Point estimate plus (+) the Margin of error is the upper bound of the confidence interval, and the Point estimate minus (−) the Margin of error is the lower bound of the confidence interval. Consequently, the size of the margin of error determines the size of the confidence interval. A larger margin of error creates a larger confidence interval to help ensure that the interval contains the true population parameter. Or,

$$\text{Lower bound} < \text{Population parameter} < \text{Upper bound}$$

Let's look at how to compute a confidence interval to estimate the population mean based on a sample. To calculate the confidence interval, we need:

1. the sample mean ($\bar{x}$)
2. the sample standard deviation (s)
3. the sample size (n), and
4. the t multiplier ($t_{\alpha/2,\, df}$)

We combine these items to get the confidence interval using this formula:

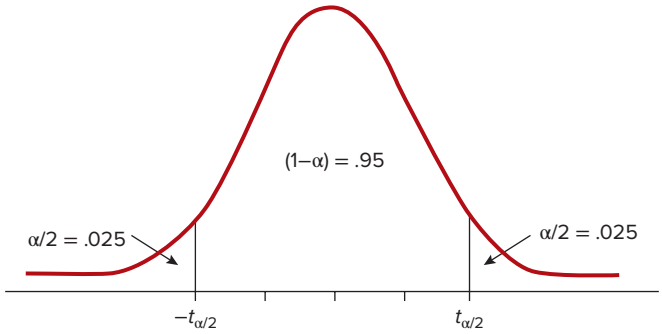
$$\bar{x} \pm \underbrace{t_{\alpha/2,\, df}\frac{s}{\sqrt{n-1}}}_{\text{Margin of Error}}$$

$$\text{Confidence Interval} = \text{Point Estimate} \pm \text{Margin of Error}$$

[1]This section is based on Statistics Online: S.2 Confidence Intervals, available at: https://online.stat.psu.edu/statprogram/reviews/statistical-concepts/confidence-intervals. Accessed 08/19/2022.

We have defined all of the terms except the *t* multiplier ($t_{\alpha/2,}$ *df*), which is a function of:

1. the degrees of freedom (*df*), the number of values that can vary in the calculation, typically measured as sample size *n* minus 1, and
2. the significance level, or alpha (α), which is set by the business analyst at the beginning of the analysis and indicates the probability that the actual population value is outside the confidence interval and is in the tails of the *t* distribution. The figure below shows the *t* distribution with α equal to 0.05, making each tail (α/2) have a probability of 0.025. In the middle of the *t* distribution is the confidence level, or 1 − alpha (α), equal to a 95 percent chance that the true population mean is in the confidence interval.



The value of the *t* multiplier is typically looked up using the *t*-distribution table found in statistics texts, reference books, or online.

Sample *t*-table

The *t* multiplier is a function of the alpha (α) and the degrees of freedom ($n - 1$). For example, (1) assume you want a 95 percent confidence level, or alpha equal to 0.05, and (2) degrees of freedom equals 4. Because the confidence interval has both left and right tails, you would use the two-tailed probability of 0.05 where it intersects the degrees of freedom of 4, or 2.7764.

One Tail	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
Two Tails	0.1	0.05	0.02	0.01	0.005	0.002	0.001
df							
1	6.3138	12.7065	31.8193	63.6551	127.3447	318.4930	636.0450
2	2.9200	4.3026	6.9646	9.9247	14.0887	22.3276	31.5989
3	2.3534	3.1824	4.5407	5.8408	7.4534	10.2145	12.9242
4	2.1319	2.7764	3.7470	4.6041	5.5976	7.1732	8.6103
40	1.6839	2.0211	2.4233	2.7045	2.9712	3.3069	3.5510
41	1.6829	2.0196	2.4208	2.7012	2.9670	3.3013	3.5442
42	1.6820	2.0181	2.4185	2.6981	2.9630	3.2959	3.5378
43	1.6811	2.0167	2.4162	2.6951	2.9591	3.2909	3.5316
44	1.6802	2.0154	2.4142	2.6923	2.9555	3.2861	3.5258
45	1.6794	2.0141	2.4121	2.6896	2.9521	3.2815	3.5202

As the sample size becomes larger (typically $n > 30$), the distribution of the values becomes closer to the normal distribution and then we use the *z* multiplier, $Z_{\alpha/2.}$ The *z* multiplier works the same way as the *t* multiplier but has the following values set values that remain constant.

Confidence Level ($1 - \alpha$)	Z Multiplier ($Z_{\alpha/2}$)
90%	1.645
95%	1.96
99%	2.58

In addition, if your data set is qualitative consisting of categories (for example, number of people wearing a blue, yellow, or green shirt), the process is the same as described for the mean but would use the proportion (p) instead of the mean ($\bar{x}$).

Calculating the Confidence Interval

Following are the formulas for computing confidence intervals.

Term	Confidence Interval Formula
Quantitative data – small sample	$\bar{x} \pm t_{\frac{\alpha}{2}, df} \frac{s}{\sqrt{n-1}}$
Quantitative data – large sample	$\bar{x} \pm Z_{\frac{\alpha}{2}, df} \frac{s}{\sqrt{n-1}}$
Qualitative data – small sample	$p \pm t_{\frac{\alpha}{2}, df} \sqrt{\frac{p(1-p)}{n}}$
Qualitative data – large sample	$p \pm Z_{\frac{\alpha}{2}, df} \sqrt{\frac{p(1-p)}{n}}$

Hypothesis Tests: Calculating the *t*-Stat

The hypothesis test in Excel calculates a t-stat. A t-test can be used to compare two groups. There are several types of t-tests. We focus on differences in means.

1. A one-sample test tests whether the mean is different than a hypothesized value (here, μ).
2. A two-sample test (also called an independent samples t-test) tests whether two groups are different. The groups may have a different number of observations.
3. A paired-sample test tests whether a sample has changed over time (that is, before group versus after group) or if two sets of observations are different over time. The groups must have the same number of observations.

The following formulas are used to calculate the t-stat in a hypothesis test, which determines if differences in measures (means) are statistically significant, or due only to chance.

Term	*t*-stat Formula
Quantitative one-sample	$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
Quantitative two-sample (independent, assuming unequal variances)	$t = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{s_a^2/n_a + s_b^2/n_b}}$
Quantitative paired-sample	$t = \frac{\bar{d}}{\sqrt{s_d/\sqrt{n}}}$ where d = difference per paired value $\bar{d}$ = mean difference of paired values s_d, = standard deviation of paired differences = $\sqrt{\frac{\sum(d-\bar{d})^2}{n-1}}$

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Installing Excel's Analysis ToolPak Add-In

Excel's Analysis ToolPak add-in allows you execute predefined analytics such as Descriptive Statistics and Regression Analysis.

Task 1: Determine if the ToolPak is installed.

1. Open Excel.
2. From the menu, select **Data > Analyze.** If Excel looks like Exhibit E1, then the ToolPak is installed, and you are done! If the Data Analysis menu option is missing, then the ToolPak needs to be installed. Follow the instructions below to install.



Exhibit E1

Task 2: Install the ToolPak.

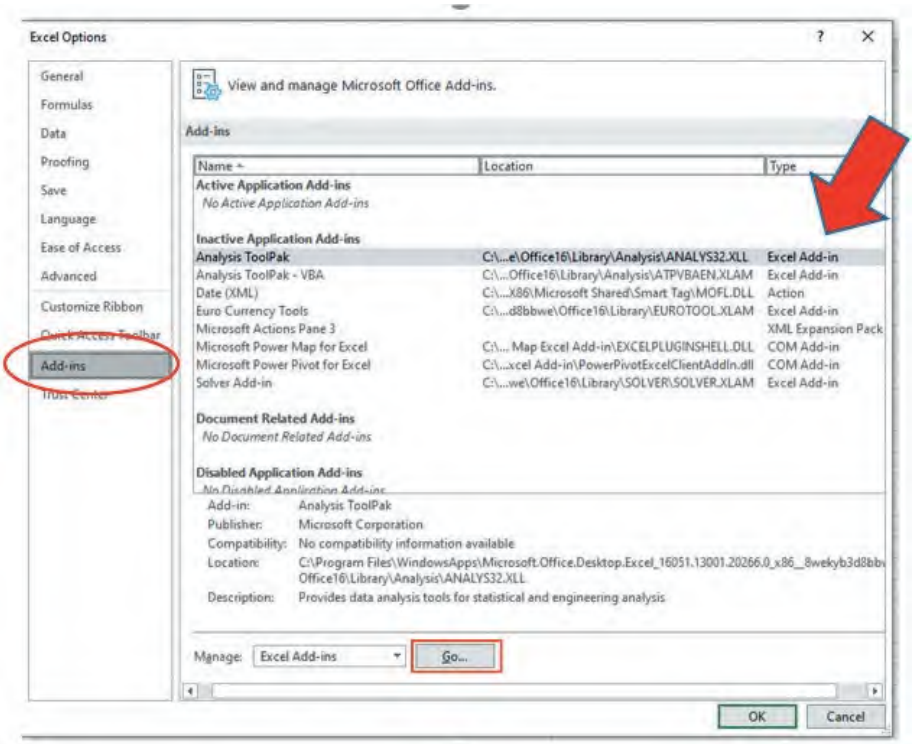
1. From the menu, select **File> Options** (at the bottom left of the screen, as shown in Exhibit E2).

Exhibit E2



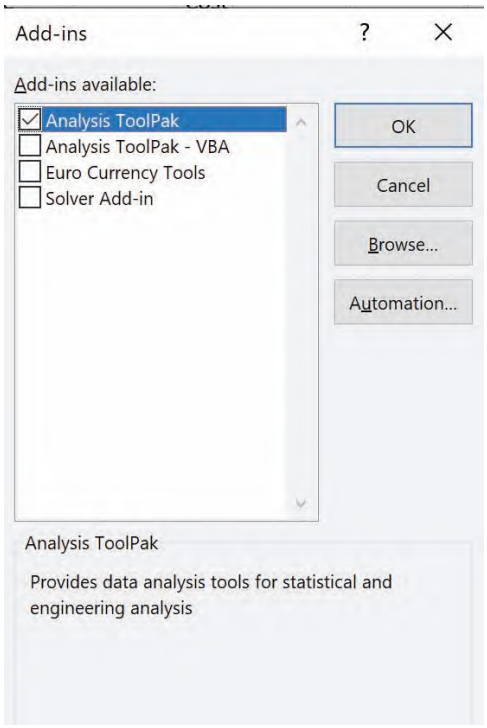
2. The Excel Options dialog box should open as shown in Exhibit E3. Select Add-ins.

Exhibit E3



3. Select the Analysis ToolPak add-in.
4. Click Go….
5. The Add-ins dialog box should appear as shown in Exhibit E4. Make sure the Analysis ToolPak add-in box is checked. Click OK.

Exhibit E4



6. From the menu, click the Data tab to see Data Analysis in the Analyze group, as shown in Exhibit E5. You are now ready to use the Analysis ToolPak.



Exhibit E5

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Installing Excel's Solver Add-In

Excel's Analysis Solver Add-In allows you to perform what-if analysis and find the optimal value for a formula. The optimal value can be a minimum or maximum depending on the question you are trying to answer

Task 1: See if the Solver is installed.

1. Open Excel.
2. From the menu, select **Data > Analyze.** If Excel looks like Exhibit F1, then the Solver is installed, and you are done! If the Solver menu option is missing, then the Solver add-in needs to be installed. Follow the instructions below to install.

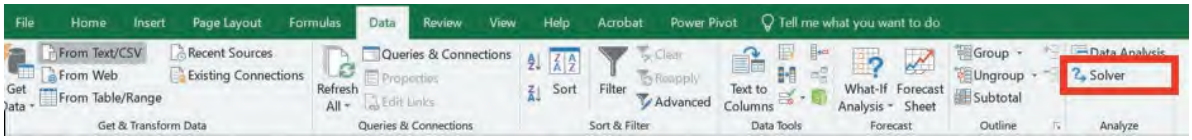


Exhibit F1

Task 2: Install the Solver add-in.

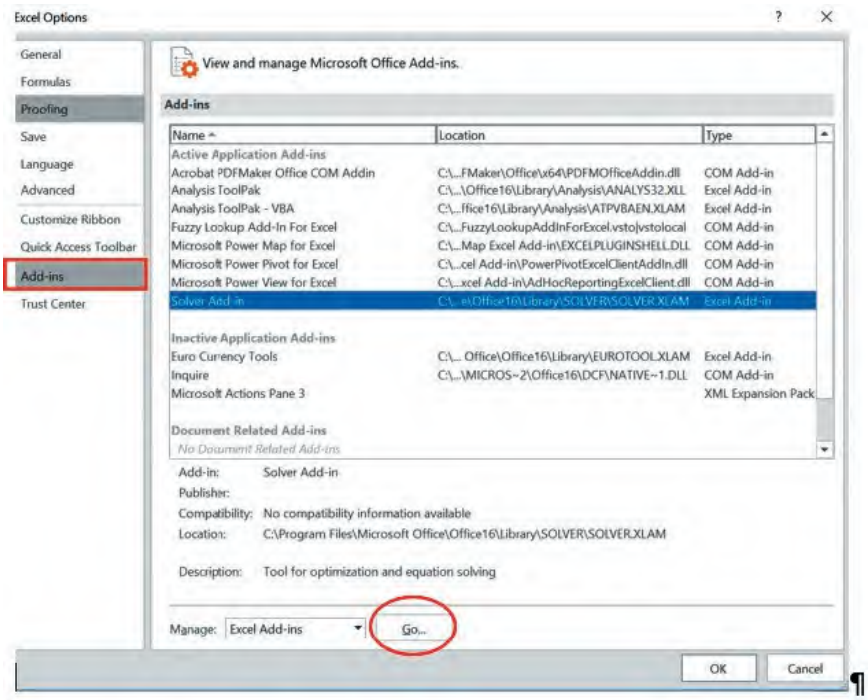
1. From the menu, select **File> Options** (at the bottom left of the screen, as shown in Exhibit F2).

Exhibit F2



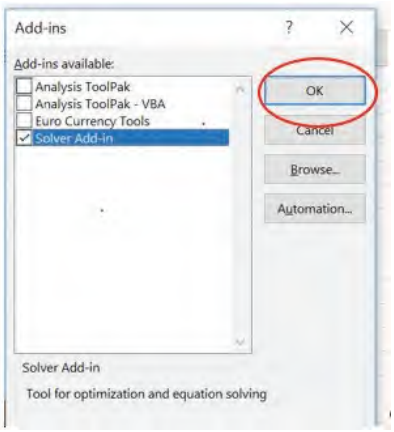
2. The Excel Options window should open as shown in Exhibit F3. Select Add-ins.

Exhibit F3



3. Select Solver Add-in.
4. Click Go….
5. The Add-ins window should appear as shown in Exhibit F4. Make sure Solver Add-in is checked. Click OK.

Exhibit F4



6. From the menu, click the Data tab to see the Solver in the Analyze group, as shown in Exhibit F5. You are now ready to use the Solver add-in.



Exhibit F5

Unless otherwise credited screenshots: Source: Microsoft Excel
Design Elements: (Vector Scales): K3Star/Shutterstock; Checkmark icon, (Briefcase icon): McGraw Hill; (Set of flat icons): PureSolution/Shutterstock; (Arrow logo): Nada design/Shutterstock; (Circular icons): miakievy/DigitalVision Vectors/Getty Images.

Index

A

Accenture, 541
Accounting, 542
Accounting analytics, 9
  descriptive analytics
    common-size financial statements, 555
    horizontal analysis, 553–554
    summarization and statistical techniques, 552
    vertical analysis, 554
  diagnostic analytics
    drill-down analysis, 559–560
    fuzzy matching, 558
    identifying anomalies and outliers, 556–559
    sequence check, 557–558
  obtain the data
    auditing data sources, 551
    financial accounting data sources, 545–548
    managerial accounting data sources, 548–551
    tax data sources, 551
  predictive analytics, 561–562
  prescriptive analytics
    goal-seek analysis, 563
    sensitivity analysis, 564
  report the results
    graph, 567
    tables, 565–566
    visualizations, 566
  role of accounting
    auditing, 542
    financial accounting, 542
    generally accepted accounting principles (GAAP), 542
    internal auditors, 542
    managerial accounting, 542
    tax accounting, 542
  specifying the question, 543–544
  time series analysis, 561–562
Adaptive/autonomous analytics, 15, 223, 777
  accounting analytics, 786–789
  analytics ascendancy model, 776
  financial analytics, 789–791
  foresight, 776
  hindsight, 776
  insight, 776
  marketing analytics, 784–786
  mimicking human capabilities, 777–780
  operations analytics, 791–798
Adjusted R square, 172
Adobe, 542
Advanced analytics methods
  artificial intelligence (AI), 780–781
  artificial neural networks, 781–783
  blockchain, 840–842
  machine learning (ML), 781
Adventure Shoes, 307, 315, 318, 323
Aerosoles.com, 688
Aggregated data, 78
AIOps platforms, 794, 795
Airbnb, 542, 785, 846, 847
  analyze the data, 850
  obtain the data, 847–849
  report the results, 851
  specify the question, 847
Air Force Materiel Command, 698
Alibaba, 277
Alphabet, 629
Alternative hypothesis, 158
Alteryx, 90
Altman, Edward, 634, 635, 671, 673, 674, 676, 677
Amazon, 3, 4, 8, 23, 82, 349, 350, 360, 441, 553, 627, 635, 644, 693, 728, 745, 797, 828, 857
Amazon Web Services (AWS), 76
Ambiance Apparel, 790
American Airlines, 8, 466
Analysis of variance (ANOVA) test, 165
Analytics ascendency model, 15, 15f
Analytics mindset, 5, 9
Anheuser-Busch, 337, 688
Anomalies, 627
Anscombe, Francis J., 339
Apple Inc., 3, 23, 278, 542, 546, 553, 554, 577, 582, 624, 634, 635
Apple Watch Series 7, 6
Artificial intelligence (AI), 780
Artificial neural networks (ANNs), 782
Asset turnover, 628, 701
AT&T Inc., 705
Auditing, 542

B

Baa corporate bonds, 626, 632
Balance sheet, 227, 546
BambooHR, 226
Bank of America, 4, 620
Bank reconciliations
  bank service fees, 235
  electronic funds transfer, 235
  errors, 235
  non-sufficient funds (NSF) checks, 235
  notes and interest, 235
  outstanding checks, 234
  outstanding deposits, 234
Barbie, 439
Bar charts *vs.* histograms
  aged receivables, 353
  bins, 351
  class intervals, 351
  count of observations, 352
  customer wait time, 352
  descriptive statistics, 352
  numerical data, 352
Barilliance E-commerce analytics, 455
Base rate fallacy, 280
Basic statistics and tools
  analysis of variance (ANOVA) test, 165–166
  bias in business analytics
    confirmation bias, 140
    nonresponse bias, 140
    outlier bias, 140
    selection bias, 140
  chi-square test, 166
  correlation, 168–169
  data reduction, 138–139
  decision process, 139–140
  descriptive/summary statistics, 137
  discrete data distributions, 174
  hypothesis testing
    alternative hypothesis, 158
    confidence interval, 157
    confidence level, 157
    confirmatory analysis, 157
    exploratory analysis, 156
    hypothesis tests, 157
    margin of error, 157
    null hypothesis, 158
    one-tailed hypothesis test, 158
    point estimate, 157
    p-values, 161–164
    set statistical significance, 158–159
    standard error, 157
    test statistic, 159–160
    two-tailed hypothesis test, 158
  inferential statistics, 137
  interpreting and visualizing statistics
    box plots, 155–156
    descriptive statistics, 153

interpreting and visualizing statistics—*Cont.*
frequency distribution, 154
histograms, 154–155
parameters and statistics, 136–137
population and samples, 136
regression analysis
dependent variable, 170
independent variable, 170
line of best fit, 171
multiple regression analysis, 172
purpose and interpretation, 173
regression coefficients, 171
R Square value, 171, 172
SAT scores, 170
Significance F test, 171
sampling methods
cluster sampling, 138
convenience/non-probability sampling, 138
simple random sampling, 137, 138
stratified random sampling, 138
software tools
Data Analysis ToolPak, Excel, 149–150
default aggregate data, 150–153
Excel functions, 148–149
tableau summary desktop card, 152
statistics
continuous probability distributions, 146–147
data distribution, 141
measures of central tendency, 142–144
measures of dispersion, 144–145
probability distribution, 141–142
random variables, 141
subset of the data, 139–140
t-tests, 164–165
Beane, Billy, 239
Benford's law, 235–238
Best Buy, 276
Bias, 140
Big Data, 74–76
data lake, 76
data warehouse, 76
distributed computing, 76
OLAP, 76
variety, 75–76
velocity, 76
veracity, 76
volume, 74–75
Binance Coin, 802
Binomial distribution, 174
Bitcoin, 802
BKD, 542
Blessing, Irish, 776
Blockchain analytics, 801–805
cryptocurrencies, 802
immutability, 803
miners, 803
proof-of-work, 803
smart contracts, 803
Bloomberg, 622, 623
BMW, 804
Boston Consulting Group, 4
botkeeper.com, 789
Bounce rate, 456
Box-and-whisker plots, 155–156, 347
Box plot, 155–156
Brand, Peter, 239
Brin, Sergey, 688
Budget, 549
Burger King, 449
Business analysts, 4, 12
Business analytics to address business questions
analytics mindset, 9
business analyst role, 7–8
businesses create value, 3–5
data *vs.* information, 6
explanatory visualizations, 18–21
exploratory visualization, 17–18
financial analytics, 8
marketing analytics, 8
operations analytics, 9
SOAR analytics model, 9–17
analyze the data, 11, 14–16
obtain the data, 11–13
report the results, 11, 16
specify the question, 11–12
Business data sources
Big Data
data lake, 76
data warehouse, 76
distributed computing, 76
OLAP, 76
variety, 75–76
velocity, 76
veracity, 76
volume, 74–75
data ethics, 92
data reports *vs.* creating reports, 69
external data sources
publicly available financial data, 74
small business administration (SBA) data, 73–74
social media data, 72–73
U.S. Census Bureau data, 73
gathering data, 92
internal data sources
centralized databases, 71–72
ERP system, 71
relational database, 71, 72
obtaining the data
tabular data, 77–78
text data, 77
preparing data for analysis
cleanse the data, 86–89
data completeness, 85–86
data integrity, 85–86
data quality, 84–85
preliminary exploratory analysis, 89
structured query language (SQL), 86
protecting data, 92
software tools
Alteryx, 90
Excel, 90
Gretl, 91
OpenRefine, 90–91
Power BI, 90
Power Query, 90
R and Python, 91
Tableau, 90
Tableau Prep, 90
structured data types
categorical data, 78–81
data dictionary, 83
date data type, 83
dimension, 83
extract, transform, and load (ETL), 84
geographic data, 83
measure, 83
number data type, 83
numerical data, 81–82
string, text, short text, or alphanumeric data type, 83
Business process, 4
Business value, 4

C

Cairo, Alberto, 361
Calcbench, 546
CalPERS, 621
Capital asset pricing model (CAPM), 633
Capital budgeting, 642
Cardana, 802
Carvana, 440
Cash flow analysis, 282
Cash-to-cash cycle, 701
Categorical data, 78, 342
bar charts, 343
heat maps, 346, 347
nominal data, 79–80
ordinal data, 80–81
pie charts, 343
stacked bar charts, 343, 344
symbol maps, 345, 346
tree maps, 345
word clouds, 345, 346
Cause-and-effect diagram, 457
Center for Systems Science and Engineering, 845
Centralized databases, 71

Certificate of deposit (CD), 631
Charles Schwab's, 779
Chatbot, 792
Checksum, 86
Chicago Board Options Exchange (CBOE), 620
Chick-Fil-A, 166
Chipotle, 555, 622
Chi-square test, 166
Churn, 456
Cinematch movie recommendation system, 16–17
City of Chicago, 699
Clean functions, 87
Cloudera, 76
Cluster analysis, 458
Cluster LeLights', 519
Cluster sampling, 138
Cockpits, 800
Cognitive analytics, 777
Common-size financial statements, 555
Compustat, 74
Conditional formatting, 233
Confidence interval, 157
Confidence level, 157
Confirmation bias, 140
Confirmatory analytics, 225, 342, 356
  base rate, 280
  base rate fallacy, 280
  predictive analytics
    classification, 274–277
    regression analysis, 277–279
    time series analysis, 279–280
  prescriptive analytics
    cash flow analysis, 282
    changing conditions faced by the company, 281
    constraints faced by the company, 281
    goal-seek analysis, 282
    marginal/incremental analysis, 281
    questions and specific techniques employed, 283
    scenario analysis, 282
    sensitivity analysis, 283–284
    what-if analysis, 281
  probabilistic models, 274
Confirmatory data analytics, 274
Consumer Reports, 706
Context, 6
Continuous data, 142, 347
Continuous probability distributions
  normal (Gaussian) distribution, 146
  standard normal distribution, 146–147
  uniform distribution, 147
  z-scores, 146–147
Control charts, 704
Convenience/non-probability sampling, 138
Coors, 447
Corey Stanford Consulting, 642
Corporate finance, 620
Correlation, 168, 241, 357–358
CORREL function, 460
Cost behavior analysis, 559
Costco, 12
Cost of capital, 639
Cost-per-click (CPC) fees, 465, 530
Cost-per-mille (CPM) fees, 465, 530
COVID-19, 644, 845
  panic buying, 273
Credit score, 276
Cross-tabulations, 559
Crowe Horwath, 789
Cuban, Mark, 467
Customer acquisition cost, 454
Customer journey analytics, 449
Customer lifetime value, 454
Customer purchase patterns, 276
Customer relationship management (CRM), 226, 227, 444, 548, 550, 695, 696
Customer touchpoints, 275
Customs and Border Protection, 698
Cybersecurity analytics, 694

D

Daimler, 807
Damodaran, Aswath, 623, 630
Danone, 797
Data, 6
  analysts, 12
  completeness, 85
  dictionary, 83
  distribution, 141
  ethics, 92
  integrity, 12, 85
  mining, 781
  overload, 5
  reduction, 138
  scientist, 7
Data-driven organizations, 805
  future of business analytics, 806–807
  limitations of business analytics, 806
Data lakes, 76, 781
Data visualization, 17, 338
  basics of, 338–342
  chart types, 342–348
    categorical data (*See* Categorical data)
    numerical (quantitative) data (*See* numerical/quantitative data)
  confirmatory analytics
    correlation, 357–358
    forecasting with time series data, 358–359
    regression, 358
  data point, 340
  data presentations
    checklist for creating effective charts, 361–362
    dashboard for Amazon, 360
  data series, 340
  executive summary, 362–364
  exploratory business analytics, 348–356
    bar charts, 349–350
    histograms *vs.* bar charts, 351–353
    line chart, 350
    pie charts, 350–351
    Pivot Tables, 355–356
    ratio data, 350
    time series data, 349
    visualizing outliers and anomalies, 354–355
  full report, 363–364
  horizontal axis, 340
  infographics, 338
  scale, 340
  sorting considerations, 340–342
    bar chart, 341
    categorical data, 342
    nominal data, 342
    numerical data, 342
    ordinal data, 342
  tick marks, 340
  vertical axis, 340
Data warehouses, 76, 781
Decision tree, 463
Deep learning, 782
Dell, 688
Deloitte, 779
Delta, 807
Demand planning, 797
Deming, W. Edwards, 705
Dependent variables, 170
Depreciation expense, 551
Descriptive analytics, 14, 223, 226–231
  accounting analytics
    common-size financial statements, 555
    descriptive statistics, 553
    horizontal analysis, 553–554
    summarization and statistical techniques, 552
    vertical analysis, 554
  exploratory business analytics, 223
    customer relationship management (CRM), 227
    data visualizations, 229
    financial reporting systems (FRS), 227
    horizontal analysis, 230
    human resource management system (HRMS), 226
    SCM system, 227

Descriptive analytics—*Cont.*
summarization and statistical techniques, 228
using descriptive statistics, 228–229
vertical analysis, 231
marketing analytics, 451–456
Barilliance E-commerce analytics, 455
churn, 456
customer acquisition cost, 454
customer lifetime value, 454
horizontal bar chart for LEGO set releases, 453
key performance indicators (KPIs), 454
leads, 454
LEGO, 451–456
return on marketing investment (ROMI), 455
sales revenue, 454
statistical techniques, 452
summarization and statistical techniques, 451
website information, 456
year-over-year (YOY) growth, 456
operations analytics
head count, 698–699
key performance indicators (KPIs), 700–702
statistical techniques, 698
summarization and statistical techniques, 697–702
website response time, 699
Descriptive statistics, 228, 241, 352
Diagnostic analytics, 14, 223, 232–237
accounting analytics
drill-down analysis, 559–560
fuzzy matching, 558
identifying anomalies and outliers, 556–559
sequence check, 557–558
exploratory business analytics, 223
drill-down analytics, 238–239
identifying anomalies and outliers, 232–238
statistical analysis, 239–240
marketing analytics, 456–462
cluster analysis, 458–459
correlation analysis, 459–462
root cause analysis, 456–458
sales funnel chart, 460–461
operations analytics
control charts, 704–706
hypothesis testing, 703–704
identifying anomalies/outliers, 702–703
Diet Coke, 447
Digital audit, 786–788
Digital ledger, 802
Discount/interest rate, 639
Discrete data, 142, 347
DISH, 804
Disney, 555
Dogecoin, 802
Domino's Pizza, 166, 690, 692
Don't Fall Inc., 758
DoorDash, 4, 542
DOT, 802
Doyle, Arthur Conan, 456
Drill-down analytics, 232, 238–239, 457
Drilling through, 457
Drilling up, 457
Driverless driving, 775
DuckDuckGo, 456
DuPont ratio analysis, 628
Dynamic pricing, 785
Dynamic report, 16

E

Eagle Golf Company, 768, 773, 821, 827, 836
Earnings call transcript, 549
Eat-it-Fast, 831, 835
eBay, 4
Electronic Data Gathering, Analysis, and Retrieval System (EDGAR), 624
Electronic funds transfer (EFT), 236
Employee survival analysis, 794
Employment length, 853
Enterprise resource planning (ERP), 71, 694–696
Enterprise system, 71
Equifax, 448
Ethereum, 802
Etsy, 804
Excel, 78
activity-based costing, 211–219
aggregating and visualizing data types, 127–132
applying Benford's Law, 262–270
applying scenario analysis, 331–334
assessing the returns to education, 324–330
bank reconciliation using conditional formatting, 258–261
customer profitability, 39–43
customer repurchase rates using cohorts, 821–827
Data Analysis ToolPak, 240–242
data filteration, 865
data manipulation, 863–865
descriptive analytics
calculate KPIs for Walmart and Amazon, 728–733
company historical performance, 482–489
dashboard creation, 413–420
historical performance using Pivot Table, 490–496
horizontal analysis, financial performance of Apple, 577–581
KPIs for Bed, Bath, and Beyond and Target, 733
KPIs for Walmart and Amazon, 728–733
Pivot Table data using conditional formatting and sparklines, 377–383
vertical analysis, financial performance of Apple, 582–585
diagnostic analytics
calculate KPIs for Walmart and Amazon, 745–749
estimating fixed and variable costs, 592–597
KPI differences between Bed, Bath, and Beyond and Target, 748–749
KPI differences between Walmart and Amazon, 745–748
look for fraud, 586–591
sales process with a sales funnel chart, 507–513
evaluating inventory, 253–257
forecasting product demand
using regression, 319–323
using time series analysis, 305–307
formatting for an income statement, 860–863
identifying and working with data types, 104–108
Monte Carlo simulation, 836–839
performing an ANOVA test, 206–211
performing a *t*-test, 201–206
pivot tables, 866–867
predictive analytics
forecasting unit demand, 758–767
future performance of IBM, 598–601
optimizing employee assignment, 768–773
sales revenue from advertising expense, 520–524
prescriptive analytics
break-even point determination, 614–617
calculating internet CPM rate, 530–533
calculating product price, 534–538
optimizing employee assignment, 768–773
product (SKU) profitability, 52–57
profitability of a mowing business, 32–38

sales and R&D expenditures, 293–298
nonlinear relationship, 299–304
summary statistics calculation
using Excel Analysis ToolPak, 193–196
using Excel functions, 189–192
supply chain
KPI differences between Bed, Bath, and Beyond and Target, 748–749
KPI differences between Walmart and Amazon, 745–748
KPIs for Bed, Bath, and Beyond and Target, 733
KPIs for Walmart and Amazon, 728–733
visualizing survival analysis, 831–835
Exotic Fruits, 253
Experian, 448
Explanatory data visualizations, 17, 18
Exploratory business analytics, 225, 342, 348–356
adaptive/autonomous analytics, 223
bar charts, 349–350
vs. histograms, 351–353
definition, 225
descriptive analytics, 223
CRM system, 227
data visualizations, 229
financial reporting systems (FRS), 227
horizontal analysis, 230
human resource management system (HRMS), 226
SCM system, 227
summarization and statistical techniques, 228
using descriptive statistics, 228–229
vertical analysis, 231
diagnostic analytics, 223
drill-down analytics, 238–239
identifying anomalies and outliers, 232–238
statistical analysis, 239–240
line chart, 350
pie charts, 350–351
pivot tables, 355–356
predictive analytics, 223
prescriptive analytics, 223
ratio data, 350
SOAR analytics model, 222
specific business questions, 223–225
time series data, 349
visualizing outliers and anomalies, 355–356
Exploratory data visualizations, 17–18
External data sources, 72–74
External marketing data
culture, laws, and the business environment, 447–448
for-profit organizations data, 448
social media data, 446–447
weather data, 447
Extract, Load, and Transform (ELT), 84
Exxon, 688
EY, 5, 9, 23, 688, 804
Eyeballing Class Exam Outliers, 354

F

Facebook, 73, 337, 446, 542, 857
FactSet, 623
Fancy Fruits, 39
FANUC, 797
FastnFurious Company, 750
FedEx, 779
Fidelity, 221
Fields, 71
Filled (geographic) maps, 347
Fill rate, 697
Filtering, 138–139
Finance, 8, 620
Financial accounting data sources
balance sheet, 546
conference call transcripts, 548
financial statements, 546
Form 8-K, 548
Form 10-K, 547
Form 10-Q, 548
general journal, 545
general ledger, 546
income statement, 546, 547
press releases, 548
Securities and Exchange Commission (SEC), 546
statement of cash flows, 546
statement of stockholders' equity, 546
Financial analytics, 8
Financial capital, 620
Financial institutions, 620–621
Financial leverage, 628
Financial markets, 620
Financial reporting (accounting) systems (FRS), 226, 227
Financial statements, 227, 546
Fintech (financial technology) revolution, 791
Fiorina, Carly, 441
Fishbone, 457
Fisher-Price, 439
Fitness Guru, 830
Five Guys Burgers and Fries, 779
Fizzy, 805
Forbes, 688
Ford, 779
Ford Motor Company, 564, 705–706
Ford Taurus, 705
Foreign key, 72
Foresight, 776
Form 8-K, 548
Form 10-K, 547
Form 10-Q, 548
Fraud, 276
Frequency distribution, 154
The Fruity Fruits Company, 189, 193, 197, 201
Fulton Homes, 4
Fuzzy matching, 558

G

Gartner, 794
Gates, Bill, 440
Gathering data, 92
General journal, 545
General ledger, 234, 546
Generally accepted accounting principles (GAAP), 542, 551
General Motors, 706
Genesis block, 803, 840
Goal-seek analysis, 282
Goldman Sachs, 621
Google, 444, 445, 456, 460, 542, 581, 585, 687, 688, 775, 780, 807
Google Analytics, 72, 73, 445, 513
Google Sheets, 78
Google's human resource Zingers, 687
Google Trends, 468
Grade point averages (GPAs), 687
Gretl, 91
Grisham, John, 797
Gross margin, 238

H

Hadoop, 76
Halloween Orders, 13
Hasbro, 439
Hash, 803
Heathrow airport, 449
Heat maps, 346
Herringbone, 457
Hershey, 13, 631, 664
Hike 'Em Hiking Boots, 305, 308, 316, 319
Hindsight, 776
Histogram, 154–155, 229, 242
Home Depot, 444, 688
Honda, 11
Honda Pilot SUVs, 11
HootSuite, 72, 73
Horizontal analysis, 230
How Charts Lie, 361
HP, 441

HR analytics, 690
HR Word Cloud, 793
Hubspot, 443, 444
Human-assisted automated accounting, 789
Human-coded algorithms, 777
Human resource management system (HRMS), 226, 551, 694
Human resources (HR), 688
Hunt, J.B., 565
Husqvarna, 619, 644
Hyperledger, 802
Hypothesis testing
  alternative hypothesis, 158
  confidence interval, 157
  confidence level, 157
  confirmatory analysis, 157
  exploratory analysis, 156
  hypothesis tests, 157
  margin of error, 157
  null hypothesis, 158
  one-tailed hypothesis test, 158
  point estimate, 157
  p-values, 161–164
  set statistical significance, 158–159
  standard error, 157
  test statistic, 159–160
  two-tailed hypothesis test, 158

I

IBM, 104, 561, 598, 602, 611, 781, 783, 786, 798, 804
IBM PC, 634
IBM Research, 783
IBM Watson, 783
Idea/ACL, 16
Identifying anomalies and outliers
  bank reconciliation, 234–235
  Benford's Law, 235–238
  conditional formatting, 233
  diagnostic accounting analytics, 556–559
  diagnostic operations analytics, 702–703
  establishing expectations, 233
  management by exception, 232–233
  potential anomalies, 234
  root cause analysis, 456–458
IHOP's burger dilemma, 445, 449
Image analytics, 800–801
iMessage, 446
Immutability, 840
Income statement, 227, 546
indeed.com, 693
Independent variables, 170
Inferential statistics, 137
Information value chain, 6
Insight, 776
Instagram, 73, 337, 446
Intel, 555
Internal auditors, 542
Internal data sources, 70–72
Internal marketing data
  CRM system, 444
  ERP/ES, 443
  IHOP's Burger Dilemma, 445
  search engine optimization (SEO), 444
  web analytics, 444
Internal rate of return (IRR), 282, 639
Internal Revenue Service (IRS), 698
International House of Pancakes (IHOP), 449
International Organization for Standardization (ISO), 85
Internet of Things (IoT), 75
Interquartile range, 145
Interval data, 81
Inventory-to-sales ratio, 701
Inventory turnover ratio, 701
Investments, 620
Ishikawa diagram, 457
IT operations analytics, 690

J

James, LeBron, 565
Jobs, Steve, 635

K

Kaggle, 847, 848
Kaput, Mike, 784
Kellogg, 631, 666
Key performance indicators (KPIs), 454, 700–702
Khoros, 439
Klipfolio, 468
KNIME, 828
KPMG, 4, 9, 237–238, 542
Kraft Heinz, 631, 666
Kuki, 792
Kurtosis, 144, 731

L

Leads, 454
Learning warehouses, 797
Lebron James, what-if analysis, and taxes, 565
LEGO, 439–441, 448–454, 459–463, 465–469, 482, 490, 497, 502, 507, 514, 520, 525, 534
Lending Club, 636, 667, 851, 857
  analyze the data, 853–854
  obtain the data, 851–853
  report the results, 855
  specify the question, 851
LeTable Inc., 489, 507, 514, 525
LexisNexis, 448
Liaison, 7
Lights-out factories, 796
Line charts, 347
Line of best fit, 171
LinkedIn, 692
Liquidity, 620
Loan paid, 853
Lower control limit (LCL), 704, 705
Lowe's, 592
Lyft, 775

M

Machine learning (ML), 781
Mackay, Harvey, 698
Maecenas, 805
Mailchimp, 444
Management by exception, 232–233
Managerial accounting data sources
  budget data, 549–550
  CRM system, 550
  human resource management system (HRMS), 551
  point-of-sale transaction data, 550
  SCM system, 550
  standard cost data, 550
Marginal/incremental analysis, 281
Margin of error, 157
Marketing, 440
  defined, 440
  marketing mix, 440–441
  place, 440
  price, 440
  product life cycle, 440
  promotion, 441
Marketing analytics, 8, 441
  descriptive analytics, 451–456
    Barilliance E-commerce analytics, 455
    churn, 456
    customer acquisition cost, 454
    customer lifetime value, 454
    horizontal bar chart for LEGO set releases, 453
    key performance indicators (KPIs), 454
    leads, 454
    LEGO, 451–456
    Return on marketing investment (ROMI), 455
    sales revenue, 454
    statistical techniques, 452
    summarization and statistical techniques, 451
    website information, 456
    year-over-year (YOY) growth, 456
  diagnostic analytics, 456–462

cluster analysis, 458–459
correlation analysis, 459–462
root cause analysis, 456–458
sales funnel chart, 460–461
marketing dashboard, 468
obtain the data, 443–451
combining internal and external data, 449–450
customer journey map, 450
external marketing data, 446–449
internal marketing data, 443–445
predictive analytics, 462–464
decision tree, 463
regression, 463–464
prescriptive analytics, 464–467
goal-seek analysis, 465–466
sky-high cash flow problem, 466–467
report the results, 468–469
specifying the question, 441–443
Marketing channels, 442
Marketing dashboard, 468
Marketing mix, 440
MarketWatch, 622
Mavericks, Dallas, 467
McDonald's, 12, 166
Mean, 142
Measures of central tendency, 142–144
kurtosis, 144
mean, 142
median, 142
mode, 142–143
skewed, 143–144
symmetrical data, 143
Measures of dispersion, 144–145
interquartile range, 145
range, 144–145
standard deviation, 145
variance, 145
Median, 142
MedRec, 802
Mercedes-Benz, 807
Merrill Lynch, 621
Michaels, 12, 447
Microsoft, 440, 555
Access, 16
Azure, 76
Excel, 90
Power BI, 90
Power Query, 90
MindBridge, 786
Miners, 803, 840
Mode, 142–143
Moisture shield, 278
Moneyball (2011), 239
Money laundering, 789–791
monster.com, 693
Monte Carlo simulation, 794–796
Morgan, JP, 807
Musk, Elon, 3
MYOB (Make Your Own Ball), 768

N

NASDAQ, 620, 622, 645
Natural language processing (NLP), 792
Nerf, 439
Netflix, 12, 16–17, 24, 601, 609, 613, 645, 780, 784
Net present value (NPV), 282, 639
Netzer, Jaime, 439
Neumann, John Von, 796
New York Stock Exchange, 620
New York Times, 705
Nielsen, 448
Nike, 440
Nominal data, 79, 342
Nonprintable characters, 86
Nonresponse bias, 140
Non-sufficient funds (NSF), 235
Normal (Gaussian) distribution, 146
Null hypothesis, 158
Numerical data, 78, 81, 342
interval data, 81
ratio data, 81–82
Numerical/quantitative data
box-and-whisker plots and histograms, 347
continuous data, 347
discrete data, 347
filled (geographic) maps, 347
line charts, 347
scatterplots, 347
NYSE, 622

O

Obtaining the data
tabular data, 77–78
text data, 77
OLAP (online analytical processing), 76
One-tailed hypothesis test, 158
OpenRefine, 90
Operations analytics, 9
descriptive analytics
head count, 698–699
key performance indicators (KPIs), 700–702
statistical techniques, 698
summarization and statistical techniques, 697–702
website response time, 699
diagnostic analytics
control charts, 704–706
hypothesis testing, 703–704
identifying anomalies/outliers, 702–703
HR analytics, 792–794
human resources (HR) process, 688
IT operations, 688
IT operations analytics, 794–796
obtain the data, 692–697
human resources data, 692–694
IT operations data, 694–695
service-level agreement (SLA), 695
service performance data, 694
supply chain data, 695–697
system performance data, 694
operations dashboard, 714–716
predictive analytics
forecasting, 708–710
regression, 708
regression analysis, 708
statistical technique, 707
prescriptive analytics
Excel solver sensitivity report, 713
optimization, 711–714
possible shipping routes, 712
statistical technique, 711
transportation optimization problem, 713
report the results, 714–716
sample HR dashboard, 715
sample IT operations dashboard, 715
sample supply chain dashboard, 716
retailers, 689
role of, 688–690
human resources (HR), 688
IT operations, 688
marketing client, 689
supply chain, 689
specify the question, 690–692
supply chain, 689
supply chain analytics, 796–798
Operations dashboard, 714
Optimization, 711
Oracle, 71, 443, 444, 695
Ordinal data, 80, 342
ORION (On-Road Integrated Optimization and Navigation) system, 797
Outdoor Adventures, 828, 829
Outlier bias, 140
Outliers, 627
Outstanding checks, 234
Outstanding deposits, 234

P

Padilla, Alex, 692
Page, Larry, 688
Parameter, 136
Patriot Software, 695
Peachtree, 443

PepsiCo, 104
Persistence, 279
Peters Engineering, 333
Petersen, Ryan, 796
Pie charts, 343, 350–351
Pinterest, 337
Point estimate, 157
Point-of-sale (POS) system, 550
Poisson distribution, 174
Population, 136
Power BI
  customer profitability, 49–51
  descriptive analytics
    dashboard creation, 430–436
    evaluate process time, 502–506
    sales revenue analysis with bar chart and filters, 405–412
    sales revenue analysis with tree map, 391–396
  desktop tutorial, 872–878
  employee turnover using a bar chart, 739–744
  forecasting product demand using time series analysis, 316–318
  future performance of IBM, 611–613
  preliminary exploratory analysis, 122–126
  preparing data types for analysis, 113–116
  product (SKU) profitability, 64–67
Predictive analytics, 14, 223, 274–280
  confirmatory business analytics
    classification, 274–277
    regression analysis, 277–279
    time series analysis, 279–280
  marketing analytics, 462–464
    decision tree, 463
    regression, 463–464
  operations analytics
    forecasting, 708–710
    regression, 708
    regression analysis, 708
    statistical technique, 707
Predictive/conditioned-based maintenance, 797
Prescriptive analytics, 14–15, 223, 281–284
  accounting analytics
    goal-seek analysis, 563
    sensitivity analysis, 564
  confirmatory business analytics
    cash flow analysis, 282
    changing conditions faced by the company, 281
    constraints faced by the company, 281
    goal-seek analysis, 282
    marginal/incremental analysis, 281
    questions and the specific techniques employed, 283
    scenario analysis, 282
    sensitivity analysis, 283–284
    what-if analysis, 281
  marketing analytics, 464–467
    goal-seek analysis, 465–466
    sky-high cash flow problem, 466–467
  operations analytics
    Excel solver sensitivity report, 713
    optimization, 711–714
    possible shipping routes, 712
    statistical technique, 711
    transportation optimization problem, 713
Press releases, 548
Price, 440
Primary key, 71
Probability distribution, 141
Procter & Gamble (P&G), 4, 6–7, 705, 800
Product life cycle, 440
Profit margin, 628
Project Gutenberg books, 783
Promotion, 441
Proof-of-work, 803, 840
Protecting data, 92
p-value, 161
PwC, 543, 781, 784
Python, 16, 91

Q

Quandl, 623
QuickBooks, 443
Quick service restaurant (QSR) Apps, 166–167

R

Range, 144–145
RapidMiner, 828
Ratio analysis, 627
Ratio data, 81, 350
Raw data, 78
R&D tax credit, 551
Rebrickable.com, 530
Rebrickable website, 451
Recommendation engines, 784
Records, 71
Regression, 242, 277, 358
Regression analysis, 170
  dependent variable, 170
  independent variable, 170
  line of best fit, 171
  multiple regression analysis, 172
  purpose and interpretation, 173
  regression coefficients, 171
  R Square value, 171, 172
  SAT scores, 170
  Significance F test, 171
Relational database, 71
  fields, 71
  foreign key, 72
  primary key, 71
  records, 71
  tables, 71
Relevant data, 12
Reliable data, 12
Research and development (R&D) expenditures, 293, 297, 299, 304
Return on equity (ROE), 629
Return on marketing investment (ROMI), 455
Reuters, 623
Richardson, Vern, 681, 684
Richardson, Vernon Jay, 586
Risk assessment, 786
Robo-advisers, 789
Robocars, 777, 778, 781
Robotic process automation (RPA), 799
Root cause analysis, 456
R programming language, 91
R software tool, 451
R Square value, 171

S

Sainsbury, 690
Sales by tax jurisdiction, 551
Salesforce, 227, 444
Sales funnel chart, 460
Sales revenue, 454
Sam's Clubs, 711
SAP, 71, 227, 443, 444
SAP Lumira, 16
SAS, 16, 451
Scatterplots, 347
Scenario analysis, 282
Schmidt, Eric, 74
Schweppes Toilet Water, 447
ScienceSoft, 804
Search engine optimization (SEO), 444
Securities and Exchange Commission (SEC), 74, 223, 546
Seeking Alpha, 548
Selection bias, 140
Semi-structured data, 75
Sensitivity analysis, 283
Sequence check, 557
Service-level agreement (SLA), 695
Service performance data, 694
Sharpe ratio, 632
Sharpe, William F., 632
Sherlock Holmes, 456
Shewhart, Walter, 705

Significance F test, 171
Simon, Bill, 13
Simple random sampling, 137
Six Sigma, 704
Skewness, 731
Sky-high cash flow problem, 466–467
Sláinte brewing company, 109, 113, 116, 117, 122, 126, 127, 132, 377, 382, 391, 395, 397, 405, 411, 413, 419, 421, 430, 434
Small Business Administration (SBA) data, 73, 74
Smart contracts, 803
SMARTRealty, 804
Smart warehouses, 797
Smucker, 631, 664
Snapchat, 337
Snapple, 447
SOAR analytics model, 10, 70, 222, 845
  Airbnb
    analyze the data, 850
    obtain the data, 847–849
    report the results, 851
    specify the question, 847
  completing your own project
    analyze the data, 856
    background and overview, 855
    deliverables, 856–858
    future aspects, 856
    obtain the data, 856
    report the results, 856
    role and motivation, 855
    specify the question, 856
  Lending Club
    analyze the data, 853–854
    obtain the data, 851–853
    report the results, 855
    specify the question, 851
Social media analytics (SMA), 446
Social media data, 72
Social Security Administration, 698
Sol, 802
Solar Roof tiles, 3
Solvency, 620
Sony, 705
Southwest Airlines, 564
S&P 500, 627, 632, 645, 649
Sparklines, 553
Speech recognition analytics, 800
Stacked bar charts, 343
Standard cost, 550
Standard deviation, 145
Standard error, 157
Standard normal distribution, 146–147
Starbucks, 166
Statement of cash flows, 546
Statement of stockholders' equity, 546
Static report, 16
Statista, 448
Statistic, 136
Stratified random sampling, 138
Structured data types
  categorical data, 78–81
  data dictionary, 83
  date data type, 83
  dimension, 83
  extract, transform, and load (ETL), 84
  geographic data, 83
  measure, 83
  number data type, 83
  numerical data, 81–82
  string, text, short text, or alphanumeric data type, 83
Structured Query Language (SQL), 86
Supply chain analytics, 689, 690
Supply chain management (SCM) system, 226, 227, 550, 695–697
Survival analysis, 792
Symbol maps, 345
System performance data, 694

T

Tableau, 90
  calculate summary statistics for company's total sales, 197–200
  customer profitability, 44–48
  descriptive analytics
    dashboard creation, 421–429, 434–436
    evaluate process time, 497–502
    sales revenue analysis with bar chart and filters, 397–404
    sales revenue analysis with tree map, 384–390
  diagnostic analytics
    employee turnover using a bar chart, 734–738, 743–744
    pricing strategy with cluster analysis, 514–519
    website response time using a control chart, 750–757
  forecasting product demand using time series analysis, 308–315
  forecasting unit demand, 758–767
  predictive analytics
    forecasting unit demand, 758–767
    future performance of IBM, 602–610
    sales revenue from advertising expense, 525–529
  preliminary exploratory analysis, 117–122
  preparing data types for analysis, 109–112
  product (SKU) profitability, 57–63
  text using a word cloud, 828, 830–831
  tutorial, 868–871
Tableau Prep, 90
Tables, 71
Taco Bell, 166
Tank, Shark, 467
Target, 629
Tax accounting, 542
Tax data sources
  tax-return data, 551
  tax-specific financial transaction data, 551
Techcrunch, 796
Telemetry system, 693–694
Term length, 853
Tesla, 3, 4, 69, 227, 565, 644, 645, 857
Tether, 802
ThrustMaster, 67, 218, 597
Tide (detergent) PODS, 6–7, 82–83
Tiff's Treats Cookie Delivery website, 445
TikTok, 446
Time series analysis, 279
Time series data, 347
Time series graph, 699
Total quality movement, 705
Toyota, 4, 23, 705
Tracr, 804
Trading Off Type I, 559
Trading Off Type II, 559
Transportation Security Administration (TSA), 698
TransUnion, 448
Tree maps, 345
Trex, 560
Trim functions, 87
TruGreen, 4
t-tests, 160, 164–165, 242
Turnover/churn rate, 785
Twitter, 73, 446, 449
Two-tailed hypothesis test, 158
Type I error/false positive, 159
Type II error or false negative, 159
Tyson Foods Inc., 230, 231, 542, 550

U

Uber, 775
UiPath, 16
Ulam, Stanislaw, 796
Uniform distribution, 147
Unilever, 779
United Way, 447
University of Arkansas, 12
Unstructured data, 75
Unusual transactions, 543
Upper control limit (UCL), 704, 705

UPS, 711, 712, 779, 797
U.S. Bureau of Labor Statistics (BLS), 5
U.S. Census Bureau data, 73
USD Coin, 802
User-specified algorithm, 462, 707
U.S. National Hurricane Center (USNC), 361
U.S. Treasury bond, 659

V

Variable, 168
Variance, 145
Velocity, 76
Veracity, 76
Vertical analysis, 231
Veterans Health Administration (VHA), 698
Voice analytics, 800–801
Voice biometrics, 800

W

Walmart, 6, 12, 13, 135, 550, 553, 627, 629, 688, 690, 711, 712, 728, 745, 748, 779, 784, 857
Walton, Sam, 688
Watson, Thomas J., 783
Wayfair, 9
Waymo, 775
Web analytics, 444
Website scraping, 799
WeChat, 446
Wells Fargo, 620, 644
Western Electric Company, 705
What-if analysis, 281
WhatsApp, 446
Wikipedia, 783
Word cloud, 345, 346, 792
Workday, 71
Working capital, 620
World Book Encyclopedia, 783
World Health Organization (WHO), 221

X

XBRL (eXtensible Business Reporting Language), 624
Xerox Corporation, 705
XRP, 802

Y

Yahoo! Finance, 69, 74, 546, 622
Year-over-Year (YOY) growth, 456

Z

Zarrella, Dan, 443
Z-scores, 146

Mc Graw Hill
501-1333